Big Data Processing and Applications

Clickstream Data for Online Shopping

Student 1: Nastaran Taefi Aghdam (2307335)
Student 2: Kamand Sedaghat Shayegan (2307265)

1 Project Description

In this project, we analyzed a clickstream dataset from an online store for maternity clothing. We started with a review of related works and a detailed description of the dataset, including schema and data types. Using Matplotlib, we visualized the data with bar and pie charts to explore distributions and relationships. Our analysis included predictive modeling with a Gradient Boosted Tree, evaluated using various techniques, yielding promising results. We also applied association rule mining to uncover significant patterns, using metrics like support, confidence, and lift. Clustering analysis revealed natural groupings in the data, enhancing our understanding. Overall, these analyses provide insights for developing further predictive models on practical features.

1.1 Motivation

Our motivation behind selecting this dataset and tools was to delve deeper into the dataset, find connections between variables, and explore potential marketing solutions. Additionally, our most important goal was to expand our knowledge of Pyspark, try out various algorithms, and train models for predictive analysis, as well as gaining good accuracy. It is worth mentioning that the dataset's rich features offered us with numerous exploration possibilities, which was the most important reason we chose it. Lastly, aiming for a top grade in our Big Data Processing Application course also made us move forward with our efforts.

2 Related work

In the field of clickstream analysis for e-shops, researchers have explored various methodologies to glean insights into user behavior and enhance the online shopping experience. Some research articles include statistical reviews and variable correlations to find the overall information and patterns of the data. [3][4] Machine learning (ML) techniques have been widely employed to predict user preferences, anticipate future actions, and personalized recommendations based on clickstream data. Moreover, predictive modeling has been leveraged to forecast purchase likelihood, enabling proactive marketing interventions and inventory management. [11][2][9] Association rule mining has been utilized to uncover patterns in user navigation paths, identifying frequently co-occurring pages or products to facilitate cross-selling and upselling strategies. [6][7][10] Clustering algorithms have also found application in segmenting users based on their browsing behavior, allowing for targeted marketing campaigns and tailored product offerings. [3][4] Overall, the integration of ML, association rule mining, prediction, and clustering techniques has enabled e-shops to utilize the information embedded in clickstream data to optimize marketing strategies, enhance user experiences, and drive business growth. A summary of related works is given in table 1.

Reference	Objective	Target	Method	Feature

Dennis Koehn et al. [2]	predict online shopping behavior and target marketing interventions in real-time	to save huge amounts of marketing costs and raise shop revenue	RNN-based classifiers	user's behavior during a session, values and averages of items viewed or put into the basket, the number of page views, the total session duration
Virgilijus Sakalauskas & Ashish Kumawat [3]	measure customer engagement and recognizing high-value customers	Targeting consumer interest patterns, and the effectiveness of advertising campaigns	compute a Customer Merit (CM) index that measures the customer's level of engagement	the customer's activity level, efficiency in selecting items, and time spent in browsing on the website and frequency of visits to e-shop
Animesh Jain & Ashish Kumawat [4]	statistical analysis to discover characteristics that influence online buying behavior	evaluating consumer behavior	statistical analysis and K-means clustering	categories of viewed products, country, product color, location, etc.
Grażyna Suchacka & Grzegorz Chodak [6]	assess a purchase probability in a user session	e-customer behavior characterization based on Web server log data.	association rule mining	categories of viewed products, user sessions with the number of session features
Orit Raphaeli et la. [7]	investigate and compare online consumer behavior on an e-retailer website in mobile versus PC devices	To find patterns with the use of a specific device and with the likelihood of a purchase.	sequential association rule mining.	session duration, average page duration, number of pages
Fernandes et la. [9]	customer purchasing engagement and predicting purchasing likelihood	in-depth understanding of online behavior	Logistic regression and the random forests models	server web logs of each customer
C. E. Dinucă [10]	learn user behavior	improve business firms, optimizing web sites and promotional campaigns	association rule mining. FPGrowth algorithm	web log data
Sabina-Cristiana Necula [11]	investigate the influence of the time spent reading product information on consumer behavior	practical implications for e-commerce website design and marketing strategies	logistic regression, decision tree, random forest, and support vector machines	the time spent on reading product information and engaging in administrative tasks

Table 1:Related works

3 Data Description

The dataset gathered on December 8, 2019, provides significant insights into the online shopping habits of users, particularly pregnant women, on a clothing store's website. Collected over five months in 2008, it provides us with a range of data points including product categories, page interactions, geographic details, and pricing information. In this section, we aim to dig deeper into the features of the dataset.

3.1 Introduction

The dataset contains 165,474 instances and 14 features, making it a valuable resource for exploring user engagement patterns and understanding consumer preferences. Analyzing this dataset can be advantageous in terms of marketing and business. For instance, by categorizing data using classification, predicting outcomes using regression, and grouping similar data points by using clustering techniques, we can gain insights into how pregnant women's products are purchased and under what circumstances. This knowledge can help us optimize product offerings, and enhance overall e-commerce performance. The whole dataset structure with examples is included in Table 2.

Interestingly, the dataset does not contain missing values, a contributing factor since missing values can skew statistical analyses and machine learning models. Moreover, complete datasets ensure integrity and they provide a more comprehensive view of the data. Without the need to

handle missing values through imputation or other methods, the analysis process becomes more efficient, which is beneficial in saving time and resources. It is worth mentioning that Mariusz Łapczyński and S. Białowąs (2013) have used this dataset to analyze user behavior and consumer buying behaviors in different countries. [1]

The dataset is licensed under the GNU General Public License Version 3, which was established by the Free Software Foundation in 2007. This license permits individuals to freely copy and distribute the dataset's content without modification, emphasizing the importance of maintaining its integrity. You can access the dataset through the following link:

https://archive.ics.uci.edu/dataset/553/clickstream+data+for+online+shopping

In summary, this dataset provides the chance to delve into complicated online shopping behavior and provide a valuable understanding of optimization of sales and marketing strategies, user experience enhancement, and growth in the competitive online retail landscape.

Field name	Description	Type	Example
Year	The year that the data is recorded	INT	2008
Month	April (4) to August (8)	INT	5 - May
Day	Day number of the month	INT	10
Order	The total number of clicks during one session	INT	2
Country	The country of origin	INT	3- Belgium
Session ID	Short unique code assigned by web server	INT	7
Page 1	The main product category	INT	2- Skirts
Page 2	A unique code for each product	STR	A13
Color	Color of the product	INT	3- Blue
Location	Photo location on the page, while the screen is divided into six parts	INT	5- bottom in the middle
Model Photography	The visual description of the product	INT	1- Only face
Price	Price in US Dollars	INT	65
Price 2	Price is higher than the average price for the entire product category	INT	2- no
Page	To which page the sold product belongs	INT	5

Table 2: Data Description with Example

4 Methods and Tools

We employed a combination of statistical analysis, machine learning algorithms, and association rule mining to gain insights into user behavior and preferences. We used PySpark, which is the Python API for Apache Spark, to process and analyze our data. Additionally, we employed Matplotlib for data visualization, and occasionally Seaborn for more plotting capabilities. For machine learning tasks, we utilized MLlib, a powerful machine learning library integrated with Apache Spark.

4.1 Data Loading and Preprocessing

We loaded data from CSV files, and performed initial data cleaning, such as removing duplicates, checking if there are any missing values, and converting data types for analysis and visualization.

4.2 Statistical Analysis and Data Visualization

We conducted descriptive statistical analysis to get insights into the characteristics of the dataset. Then, we created visualizations such as pie charts, heatmaps, boxplots, and bar plots to understand patterns and relationships in the data better and find out the best way to demonstrate the data. These visualizations allowed us to present the distribution of features, correlations between attributes, and patterns within the data

4.3 Association Rule Mining

To discover interesting relationships between the items in the dataset we applied association rule mining techniques. In this step, we utilized the Frequent Pattern Growth (FP-Growth) algorithm, to identify frequent item sets in the dataset. These frequent itemsets are then used to generate association rules based on user-defined metrics such as support, confidence, or lift.

4.4 Machine Learning

For predictive analysis, we employed the gradient-boosted tree (GBT) regression model. Initially, we experimented with the Random Forest Regressor but obtained unsatisfactory results. Subsequently, we opted for the GBT regression model, which yielded improved accuracy. We evaluated the model's performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R2), and accuracy.

4.5 Clustering Analysis

To segment the data based on price values and identify distinct customer groups, we utilized the K-means clustering algorithm. By clustering price values into meaningful groups, we gained insights into customer behavior and preferences, enabling the formulation of targeted pricing strategies and personalized recommendations.

5 Data analysis

In this section, we delve into data analysis through a series of pie charts, each offering insights into distinct aspects of the dataset.

5.1 Distribution between features

Here, we present various analyses of the distribution of values across different features by utilizing pie charts and bar charts.

5.1.1 Distribution of PRICE 2

In our analysis of attribute *price 2,* indicating whether a product's price exceeds the average for its category, we found a nearly equal distribution among our dataset. The resulting pie chart can be seen in Figure 1.

5.1.2 Distribution of LOCATION of the photo in the webpage

In our analysis of webpage image placement, we observed a rather balanced distribution across six sections, with the top left being the most significant location for images at 20.9% and the bottom right the least represented. The resulting pie chart can be seen in Figure 2.

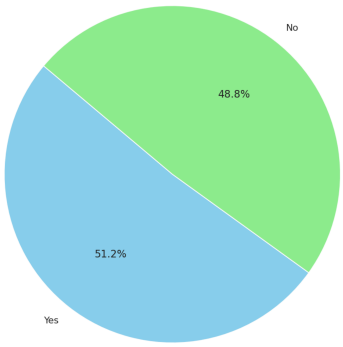


Figure 1: Distribution of PRICE 2

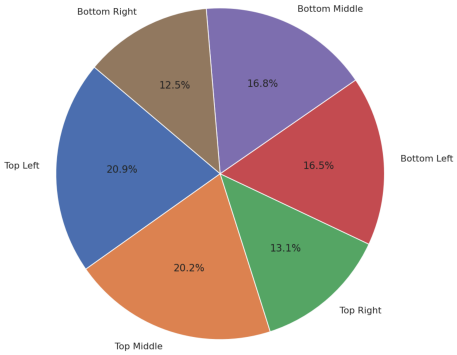


Figure 2: Distribution of LOCATION

5.1.3 Distribution of orders by COUNTRY

In the distribution of orders by country, Poland ranks highest in frequency, followed by the Czech Republic and Lithuania, as indicated by the bar chart analysis in Figure 3.

5.1.4 Distribution of orders by color

In the distribution of orders by color, black and blue rank as the most purchased colors, with burgundy being the least bought, as shown in the bar chart in Figure 4.

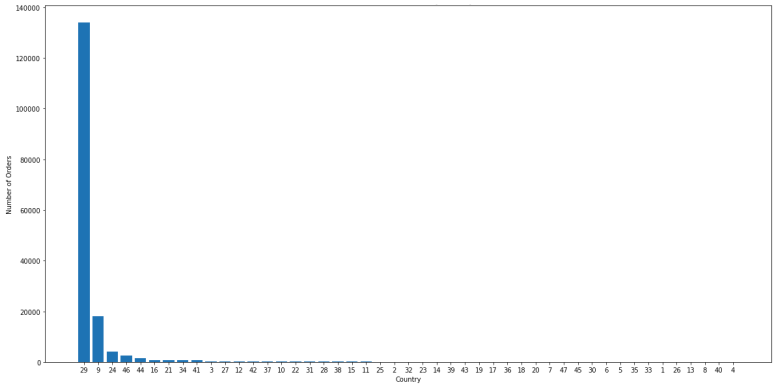


Figure 3: Distribution of orders by country

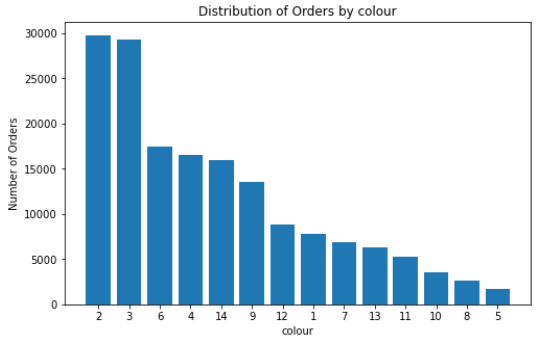


Figure 4: Distribution of orders by color

5.2 Relationship between data attributes

In this section, we explore the relationship between different data attributes to uncover correlations and dependencies within the dataset.

5.2.1 Correlation of features

In this section, we delve into the correlation of features within our dataset, as depicted in the heatmap in Figure 5. We observe a strong positive correlation between Session ID and month, while Model Photography and price demonstrate a negative correlation. Moreover, the most negative correlation belongs to “price” and “price 2”.

5.2.2 Percentage of Buying From a Category in 4 Price Ranges

We learned the price span is from 18 to 82 US dollars. So, we divided this range into four equal price ranges and calculated the proportion of each product category within these ranges. For instance, in the price range of 18 to 34, skirts were the least represented category, accounting for less than 10%. Conversely, in the price range of 50 to 66, skirts comprised the highest likelihood as visualized in Figure 6.

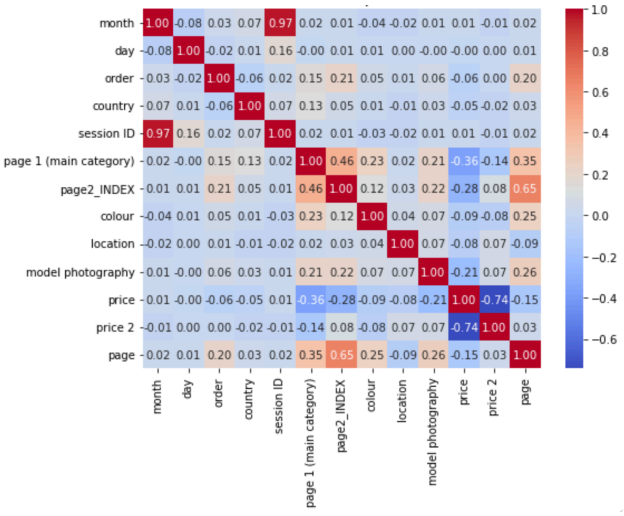


Figure 5: Correlation heatmap of features

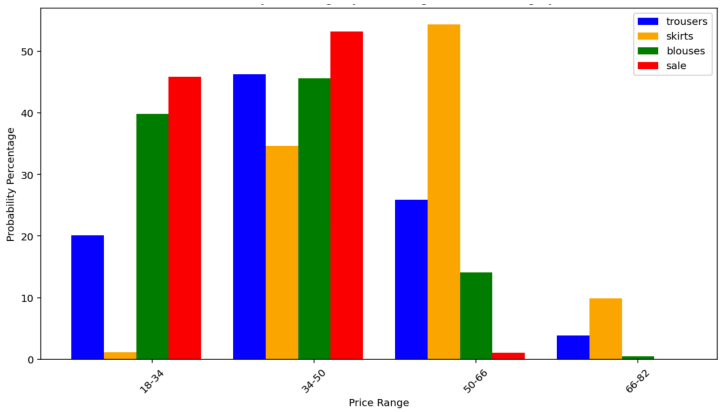


Figure 6: Probability percentage by *price range* and *main category*

5.3 Association Rules Mining, Support, Confidence, and Lift

Association rule mining reveals meaningful connections among items in the dataset, aided by metrics like confidence and lift. These concepts, when employed in market basket analysis, facilitate the discovery of item relationships in large datasets.

5.3.1 Support

Support measures how often an item appears in the dataset, indicating its popularity relative to the total transactions. It is calculated by dividing the number of transactions containing the itemset by the total number of transactions. High support values mean the item appears frequently, making it a good candidate for association rule mining.[6]

5.3.2 Confidence

Confidence measures the reliability of an association rule, indicating the likelihood that the presence of one item (the antecedent) will result in the presence of another (the consequent). It is calculated by dividing the number of transactions containing both items by the number of transactions containing the antecedent. High confidence values suggest a strong association, implying a higher probability of the consequent item being purchased when the antecedent item is present.

5.3.3 Lift

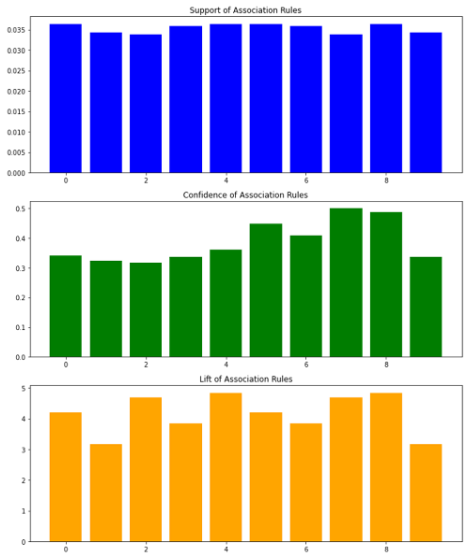
Lift measures the strength of association between the antecedent and consequent in an association rule, relative to their occurrence frequencies. It compares the observed support of the itemset with the expected support if the antecedent and consequent were independent of each other. A lift value greater than 1 indicates that the antecedent and consequent appear together more often than expected by chance, suggesting a positive association. A lift value less than 1 indicates a negative or unlikely association, while a lift value equal to 1 indicates independence.

The analysis revealed three key insights into purchasing patterns. Firstly, when item B10 is bought, the likelihood of purchasing item B13 increases by 4.85 times, indicating a strong potential for bundling. Lastly, the rule linking A3 to A2 suggests a significant increase in the

likelihood of purchasing A2 when A3 is bought. These findings offer actionable strategies to optimize marketing and enhance cross-selling tactics.

Antecedent	Consequent	Confidence	Lift	Support
[A2]	[A5]	0.34	4.21	0.0363
[A2]	[A11]	0.32	3.16	0.0343
[A2]	[A3]	0.32	4.69	0.0338
[A2]	[A1]	0.34	3.84	0.0358
[B10]	[B13]	0.36	4.85	0.0363
[A5]	[A2]	0.45	4.21	0.0363
[A1]	[A2]	0.41	3.84	0.0358
[A3]	[A2]	0.50	4.69	0.0338
[B13]	[B10]	0.49	4.85	0.0363
[A11]	[A2]	0.34	3.15	0.0343

Table 3: Association Rules For Different Products



5.4 Predictive Analysis

In our predictive analysis, we used the probability percentages derived from the past section to create a gradient-boosted tree (GBT) regression model aimed at predicting these percentages. It is worth mentioning that initially, we experimented with Random Forest Regressor, but encountered unsatisfactory results. Furthermore, we used GBT and improved our accuracy. We resulted in:

- Mean Absolute Error (MAE): 0.03974245497554473
- Mean Squared Error (MSE): 0.0064731610206736605
- R-squared (R2): 0.9999654853058938
- Accuracy: 0.89

To achieve this result, we used the GBTRegressor model with the maximum iteration of 10 and maximum depth of 5. Here is the scatter plot comparing the true labels against the predicted labels for a regression model in Figure 7.

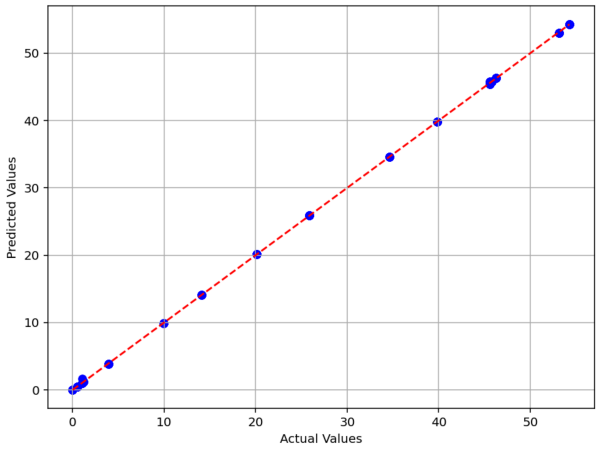


Figure 7: The actual versus the predicted values of the Gradient Boosted Tree regression model

5.5 Clustering

Online sales data often contains a wide range of price values. Clustering plays a crucial role in identifying distinct groups or segments within these price values, enabling us to group similar price points. This clustering process helps in understanding customer behavior and preferences based on their purchasing habits. By clustering price values, analysts can determine how different customer segments respond to changes in prices, thereby enabling the formulation of targeted pricing strategies aimed at maximizing revenue. Clustering analysis is a beneficial approach to investigating the price complex dynamics in online sales, facilitating data-driven decision-making for pricing strategies. Moreover, it enhances customer satisfaction by helping to identify potential customers and determining personalized recommendations based on their preferences and purchasing patterns.

To determine the optimal number of clusters, the process involves executing the K-means clustering algorithm on the dataset for various values of K, typically ranging from 1 to 10. For each value of K, the algorithm computes the Within Cluster Sum of Squares (WCSS), representing the total variations within a particular cluster. Subsequently, a curve is plotted, mapping the calculated WCSS values against the number of clusters. The point on the plot where a bend or elbow-like structure appears signifies a critical juncture. This inflection point is regarded as the optimal value of K for the clustering algorithm.[4] here, it will be concluded from Figure 8 that 3 clusters will be the optimal number where both the number of clusters and WCSS values are minimized beneficially.

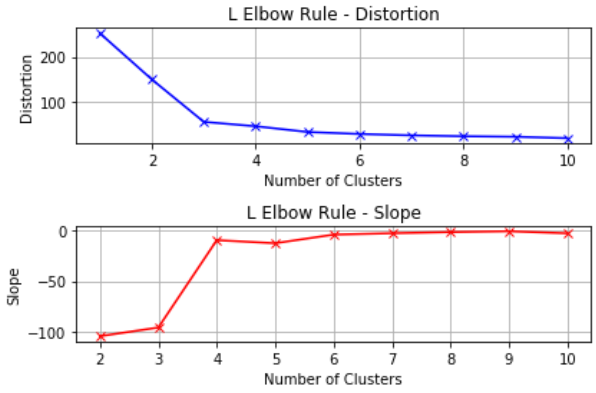


Figure 8: L Elbow rule

After clustering the data based on the price feature into three distinct clusters, each cluster reveals insightful characteristics:

Cluster 0: With a feature price of $33.12, this cluster represents lower-priced items.
Cluster 1: Featuring a price of $62.25, Cluster 1 represents higher-priced products.
Cluster 2: Positioned at a price point of $46.51, Cluster 2 signifies items with average pricing.

Once the data is segmented, each feature within these clusters can be thoroughly analyzed. For example, examining each price level allows us to identify the demand levels across various product categories represented by different colors shown in Figure 9 and Figure 10. Additionally, we can delve into the color preferences of different countries and estimate the number of interested customers per color, segmented by country. Such insights aid in pricing strategies for each product category and color preference across different regions.
The evaluation of the clustering algorithm's performance is crucial. The Silhouette Score serves as a metric to gauge the quality of the clusters formed. With a Silhouette Score of 0.594, indicating a well-separated cluster structure and cohesive objects within each cluster, the clustering algorithm demonstrates strong performance. This score falls within the range of -1 to 1, where higher values signify better-defined clusters. Hence, the obtained Silhouette Score reinforces the effectiveness of the clustering algorithm in uncovering meaningful patterns within the data.[11]

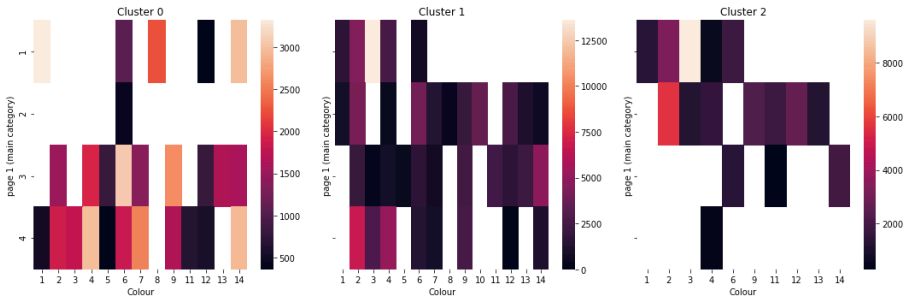


Figure 9: Order frequency in each cluster grouped by colour and main category

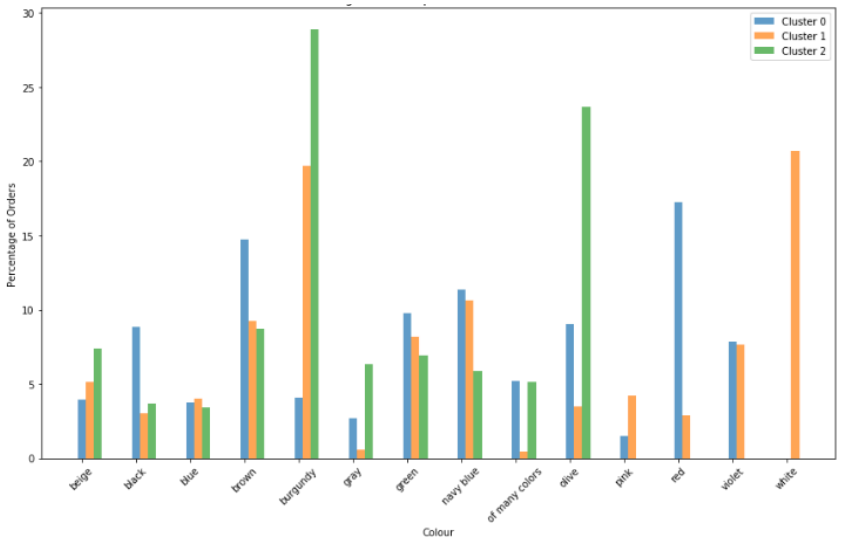


Figure 10: Order percentage in each cluster grouped by colour

6 Results

Our analysis of clickstream data from an online store for pregnant women provided valuable insights into user behavior. We used exploratory data analysis and visualization to uncover patterns. We realized that Poland ranks significantly higher in buying products, and proved that black is the most bought color. Furthermore, Predictive modeling with a gradient-boosted tree regression model accurately predicted purchase percentages across categories, outperforming a Random Forest regressor. We were able to achieve 89% accuracy with GBT, with an R-squared value of 0.99. Moreover, association rule mining revealed significant item relationships for targeted marketing. We found that the observations regarding the rules like [B10] -> [B13] and [B13] -> [B10] are significant due to their strong associations and high lift values. Lastly, clustering analysis identified customer segments based on color, which can be helpful with pricing strategies. The analysis showed that customers are more likely to buy higher prices which has a discount on the lower price level than buying average products on lower prices. Overall, the analyses provided a wide range of methods to get actionable insights for optimizing marketing strategies and competing in the online retail landscape in terms of business. In comparison with other research, this report contains a combination of various methods which will obtain a better understanding of the complex sales and customer behavior pattern. It will help businesses to provide a more flexible strategy for marketing and production, not only considering pricing policies but also the design of products.

7 Conclusion

In conclusion, looking at the clickstream data shows how important it is to use advanced analytics to understand big online shopping datasets. We used analytical techniques like statistics, machine learning, and association rules mining to learn a lot about how users behave and what they like. This knowledge can help businesses make better decisions, like creating

advertisements targeting specific groups, improving the products they sell, and making customers more satisfied. Moving forward, continued exploration and refinement of analytical approaches will be essential for staying competitive in the dynamic online retail market**.**

8 Contribution Report

We scheduled weekend meetings at the university to write code, analyze data, and explore different parts of our project. Additionally, we used a shared Google Docs file to review and improve the report, usually working on it separately. This method helped us collaborate smoothly for the project, and make sure the report was well put.

9 References

[1] Mariusz Łapczyński and S. Białowąs 2013, "Discovering Patterns of Users' Behaviour in an E-shop - Comparison of Consumer Buying Behaviors in Poland and Other European Countries" Studia Ekonomiczne

[2] Koehn, Dennis, Stefan Lessmann, and Markus Schaal. "Predicting online shopping behaviour from clickstream data using deep learning." Expert Systems with Applications 150 (2020): 113342.

[3] Sakalauskas, V.; Kriksciuniene, D. Personalized Advertising in E-Commerce: Using Clickstream Data to Target High-Value Customers. Algorithms 2024, 17, 27. https://doi.org/10.3390/a17010027

[4] Jain, Animesh, and Ashish Kumawat. "ANALYSIS OF CLICKSTREAM DATA." (2022).

[5] Yilmazcan Ozyurt, Tobias Hatt, Ce Zhang, and Stefan Feuerriegel. 2022. A Deep Markov Model for Clickstream Analytics in Online Shopping. In Proceedings of the ACM Web Conference 2022 (WWW '22). Association for Computing Machinery, New York, NY, USA, 3071–3081. https://doi.org/10.1145/3485447.3512027

[6] Suchacka, Grażyna, and Grzegorz Chodak. "Using association rules to assess purchase probability in online stores." Information Systems and e-Business Management 15 (2017): 751-780.

[7] Raphaeli, Orit, Anat Goldstein, and Lior Fink. "Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach." Electronic commerce research and applications 26 (2017): 1-12.

[8] Triandini, Evi, I. Gede Suardika, and I. Ketut Putu Suniantara. "Database Click Stream of E-commerce Functional." MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer 21.1 (2021): 75-86.

[9] Fernandes, Ricardo Filipe, and Costa Magalhães Teixeira. "Using clickstream data to analyze online purchase intentions." (2015).

[10] Dinucă, C. E. "An application for clickstream analysis." (2011). INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS Issue 1, Volume 6, 2012 68

[11] Necula SC. Exploring the Impact of Time Spent Reading Product Information on E-Commerce Websites: A Machine Learning Approach to Analyze Consumer Behavior. Behav

Sci (Basel). 2023 May 23;13(6):439. doi: 10.3390/bs13060439. PMID: 37366691; PMCID: PMC10294865.