## Introduction

In today’s digital age, e-commerce platforms are inundated with vast amounts of user interaction data, commonly referred to as clickstream data. This data, which captures the sequence of clicks and interactions a user makes while navigating a website, provides invaluable insights into user behavior and preferences. Leveraging this data to build a robust product recommendation system can significantly enhance user experience and drive sales.

The objective of this report is to explore the development of a system that recommends products to users based on their historical clickstream data. By analyzing past interactions, the system aims to predict and suggest products that align with the user’s interests and needs. This approach not only improves the accuracy of recommendations but also personalizes the shopping experience, making it more engaging and efficient.

The report will delve into the methodologies and technologies employed in the construction of such a system, including data collection, preprocessing, and the application of machine learning algorithms. Additionally, it will discuss the challenges encountered during the development process and the strategies implemented to overcome them. Through this exploration, we aim to demonstrate the potential of clickstream data in transforming e-commerce platforms and enhancing customer satisfaction.

* 1. Project Scope

This project aims to apply necessary pre-processing to transform raw data, analyze it, and train it on appropriate algorithms. So that useful insights can be drawn to recommend products to users and thus increase the customer conversion rates and provide better experience to consumers. This helps business owners in taking future decisions in the right direction.

* 1. Aims & Objectives

The primary purpose of this work experience project is to develop a robust recommendation system that enhances user engagement and satisfaction on an online shopping platform.

By leveraging historical clickstream data, the system aims to predict and suggest products or content that align with the users' preferences and behaviors, ultimately driving higher conversion rates and improving the overall shopping experience.

## Overall Description

**2.1 Clickstream Dataset:**

2.1.1 The Dataset:

Data was provided by the Shyft Company. Shyft is a Lifestyle management and tracking application for specific health conditions like chronic ailments, pregnancy care, chronic pains and more. They offer multiple different health programs that are specifically catered for the chosen health conditions. More than 30,000 customers trust Shyft to manage their chronic and health conditions.

The collection comprises clickstream data from an online business that sells apparel for pregnant women. The data are from the five months of 2008 and contain, among

Other things, product category, placement of the photo on the website, IP address country of session origin, and product selling price in US dollars. Each column in the dataset is a product bought by a customer. The dataset has 165474 rows of data with 14 variables. It has been collected by all relevant data protection laws.

2.1.2 Size of the dataset:

The data set has 165474 rows and 14 columns.

2.1.3 Attributes of the dataset:

1. **YEAR**: The data is recorded only for certain months of the same year so the value of the YEAR column is the same in all the rows i.e., 2008.

2. **MONTH**: This variable contains the month number in which the product was bought on the platform. It has values ranging from 4(April) to 8(August). This is a

Categorical variable.

3. **DAY**: This variable has the day on which the product was bought by the customer. It ranges from 1 to 31 for the number of days in a month. This is an ordinal

Variable.

4. **ORDER**: This variable contains the total number of clicks during one session of a user on the platform. It is numerical data.

5. **COUNTRY**: Variable indicating the country of origin of the IP address with the following categories:

Australia, Austria, Belgium, British Virgin Islands, The Cayman Islands, Christmas Island, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Unidentified, Faroe Islands, Finland, France, Germany, Greece, Hungary, Iceland, India, Ireland, Italy, Latvia, Lithuania, Luxembourg, Mexico, Netherlands, Norway, Poland, Portugal,

Romania, Russia, San Marino, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Arab Emirates, United Kingdom, USA, biz (.biz) 44-com (.com), Int (.int) 46-net (.net), org (\*.org).

6. **SESSION ID**: Variable indicating session-id (a short record). A session ID is a unique code assigned by a web server to a specific surfer for the duration of their visit

(Session) to the website. Session IDs are stored as cookies, form fields, or Uniform Resource Locators (URLs). Some web servers simply increment a static

Number to generate the session ID. However, most of the web servers use algorithms that include more sophisticated techniques, such as tracking even the variables such as the date and time any surfers visit the website, along with other variables defined by the server administrator. Each time a web user clicks on a particular link of a website, a new session ID is assigned. Closing and reopening the browser and revisiting the site can sometimes generate a new session ID.

7. **PAGE 1 (MAIN CATEGORY)**: This variable is concerned with the main product category. It has four categories.

1-trousers

2-skirts

3-blouses

4-sale

8. **PAGE 2 (CLOTHING MODEL)**: It contains information about the product code for each product. There are a total of 217 products in the dataset. All these products

have a unique product id associated with the respected product.

9. **COLOR**: It describes the color of the product.

1- Beige

2- Black

3- Blue

4- Brown

5- Burgundy

6- Gray

7- Green

8- Navy blue

9- of many colors

10-olive

11-pink

12-red

13-violet

14-white

10. **LOCATION**: On the web page of the website there are Locations on which the picture of the product is displayed. This variable deals with the location of the Picture on the page, the webpage has been divided into six parts:

1-top left

2-top in the middle

-top right

4-bottom left

5-bottom in the middle

6-bottom right

11. **MODEL PHOTOGRAPHY**: Any product on a website has to have a visual description. This picture has a model wearing the product. This variable depicts the visual description of the two categories:

1- Only face

2- Profile

12. **PRICE**: Price of product in US dollars.

13. **PRICE 2**: This is the variable informing about the price of a particular product being higher than the average price for the entire product category.

1-yes

2-no

14. **PAGE**: A website contains several pages to display the information and products. This variable will indicate to which page the sold product belongs. This variable has values ranging from 1 to 5.

## Data Preprocessing and Cleaning:

2.2.1 Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done.

1. handling null values

If there are less null values then they can be replaced by mean values or default values. If the number of null values is significantly high then row can be dropped. There were no null values in the data.

1. Drop duplicate

There were no duplicate rows in the data to drop.

3. Dropping irrelevant columns:

Columns Year, Month and Day are dropped from data as they are irrelevant.

4. Label encoding:

To make the data understandable or in human readable form, the training data is often labeled in words. Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Label encoding was done on the clothing model column.

5. Outliers handling:

For Outlier handling IQR technique was used. Outliers were identified in the price column. There were almost 1300 outliers which were eventually dropped.

6. Standard scaler:

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance). Continuous variables were standardized.

**2.3 Feature Analysis**:

Feature analysis is done to find out the most impacting features. Following are some plots we used to extract some useful information

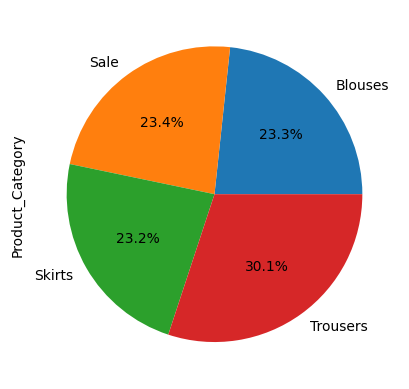


Figure 1 what kind of clothing is most clicked?

Explanation: Here we can see that Trousers have high demand whereas rest of product categories have almost same demand

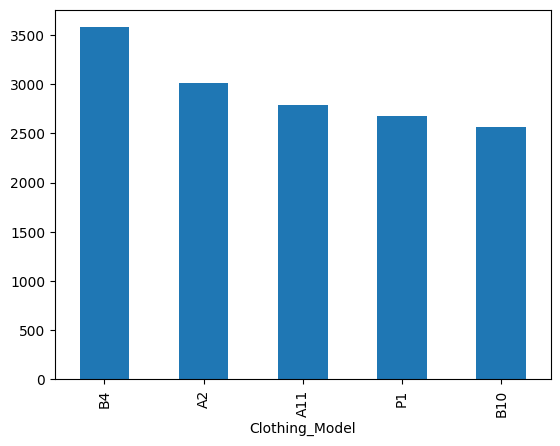


Figure 2 Top 5 clothing models.

Explanation: Here we can observe the top 5 most clicked clothing models in the dataset. This is represented using a bar plot graph. The most common model is B4.

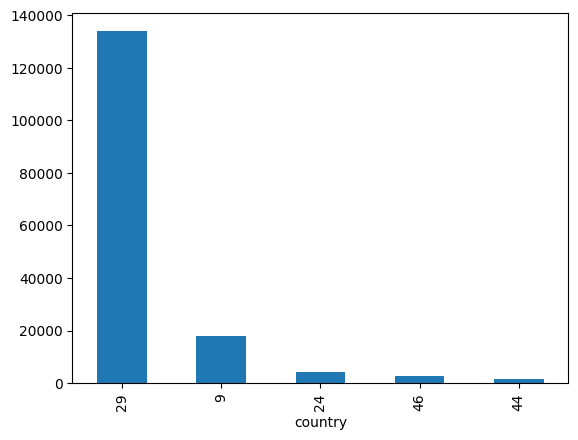


Figure 3 Top 5 country codes click wise.

Explanation: Here the top 5 countries with the most clicks are mentioned. The country with code 29 has the highest no. of clicks.

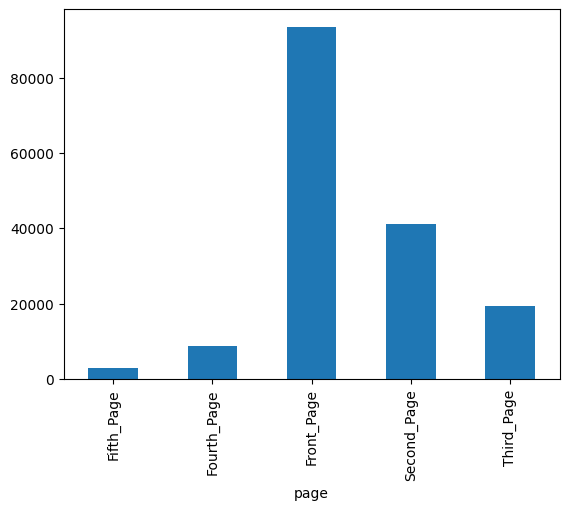


Figure 4 Table for the page with most no. of clicks.

Explanation: Here the most no. of clicks is on the Front page and the least no. of clicks is on the fifth page.

## 2.4 Customer segmentation based on session average clicks:

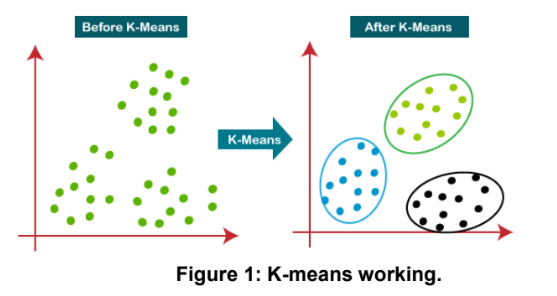


Figure 5 K means working

**K-means algorithm**

The k-means algorithm is an unsupervised machine algorithm, which makes different clusters by grouping the unlabeled dataset. The k in the K-means algorithm defines the number of predefined clusters that need tobe created while the algorithm works. It is an iterative machine learning algorithm that differentiates and divides the unlabeled dataset into k number of clusters in such a way that each data point belongs to a single group having some specific properties. It is an algorithm based on centroids, where each cluster is associated with a centroid. The main task that this algorithm does is to minimize the sum of distances between the data point and their corresponding clusters**.**

The algorithm works in the following manner: firstly, we have to select the value for K to decide the number of clusters, then the algorithm selects the random K points or centroids, then it assigns each data point to theirclosest centroid, which will form the predefined clusters, after this the variance is calculated and the new centroid is calculated and placed in each cluster, now the step of assigning the data points to the centroids is repeated. Now, the step to calculate the variance and again reassigning occurs, and the algorithm finds the defined number of clusters.

**How to choose the value of “K number of clusters” in K-means clustering?**

The efficiency of the K-means clustering algorithm depends on the formation of highly efficient clusters that it identifies and groups together. The selection of the optimal value of K is a big task. There are some ways to find out the optimal value of K for the given dataset, but the most appropriate method to find the number of clusters or K value is the Elbow Method.

**Elbow Method**

The elbow method is the most appropriate and one of the popular ways to find the optimal number of clusters. This technique used the concept of WCSS (Within Cluster Sum of Squares, which gives the total variations within a certain cluster. To measure the distance between the data points and the centroids we can use any of the methods such as Manhattan distance or Euclidean distance.

To find the optimal value of clusters, the elbow method follows the below steps:

o It executes the K-means clustering algorithm on the dataset for different values of K (in range 1-10).

o For every value of K, it calculates the WCSS value.

o Plots a curve between calculated WCSS values and the number of clusters K.

o The point of bend or a point of the plot appears to be like an arm, then that point is considered the best value of K.

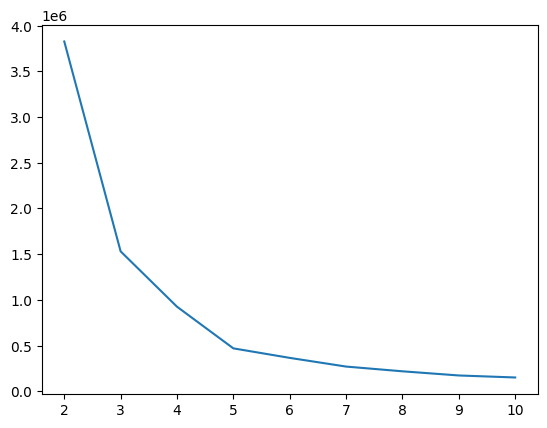


Figure 6 from elbow method k=5

## 2.5 Pattern Detection:

1. Apriori Algorithm:

The Apriori algorithm is an unsupervised machine learning algorithm used for association rule learning. Association rule learning is a data mining technique that identifies frequent patterns, connections and dependencies among different groups of items called item sets in data. Hence the Apriori algorithm was used for pattern detection in sequence of clicks.

### How the Apriori algorithm works:

Each of the key steps in the Apriori algorithm looks to identify itemsets and all their possible supersets looking for the most frequent to create the association rules.

Step 1: Frequent item sets generation

The algorithm first identifies the unique items, sometimes referred to as 1-itemsets, in the dataset along with their frequencies. Then, it combines the items that appear

together with a probability above a specified threshold into candidate item sets and filters out the infrequent item sets to reduce the compute cost in further steps. This process, known as frequent item set mining, looks just for item sets with meaningful frequencies.

Step 2: Expand and then prune item sets

Using the Apriori property, the algorithm combines frequent item sets further to form larger item sets. The larger item set combinations with a lower probability are pruned. This further reduces the search space and makes the computation more efficient.

Step 3: Repeat steps 1 and 2

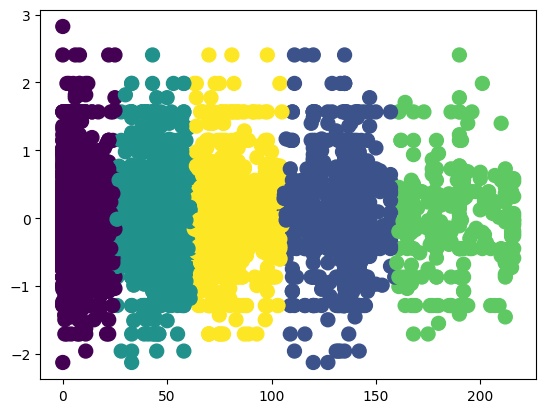


Figure 7 Customer segmentation across clusters

The algorithm repeats steps 1 and 2 until all frequent item sets meeting the defined threshold probability are generated exhaustively. Each iteration generates more complex and comprehensive associations in the item sets.

Once Apriori has created the item sets the strength of the generated associations and relationships can be investigated.

## 

## 2.6 Building Model

Train/Test split:

Data is divided into two parts: first one for training our machine learning model, and second one for testing our model.

* Split the dataset into two pieces: a training set and a testing set.
* Train the model on the training set.
* Test the model on the testing set, and evaluate how well our model did.

Advantages of train/test split:

* Models can be trained and tested on different data than the one used for training.
* Response values are known for the test dataset, hence predictions can be evaluated
* Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning models facilitate computers to find hidden insights from Big Data without being explicitly programmed where to look.

We have used the following algorithms to build a predictive model.

**Collaborative Filtering:**

Collaborative filtering uses similarities between users and items simultaneously to provide recommendations. Collaborative filtering models can recommend an item to user A based on the interests of a similar user B.

Collaborative filtering is a type of recommender system. It groups users based on similar behavior, recommending new items according to group characteristics.

Collaborative filtering is an information retrieval method that recommends items to

users based on how other users with similar preferences and behavior have interacted with that item. In other words, collaborative filtering algorithms group users based on behavior and use general group characteristics to recommend items to a target user. Collaborative recommender systems operate on the principle that similar users (behavior-wise) share similar interests and similar tastes.

**User-based filtering** recommends items to a target user based on the preferences of behaving users. The recommendation algorithm compares a target user’s past behavior to other users.

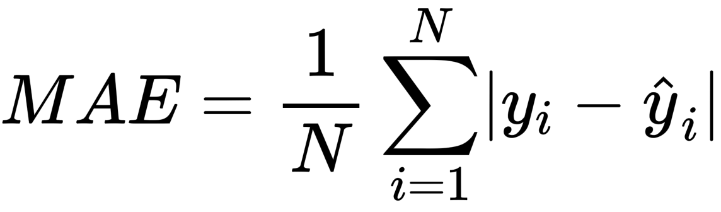
**Item-based filtering** recommends new items to a target user based on that user’s behavior toward similar items. Note, however, that in comparing items, the collaborative system does not compare item features (as in content-based filtering) but instead how users interact with those items.

**KNNBasic**

KNNBasic is a basic collaborative filtering algorithm.

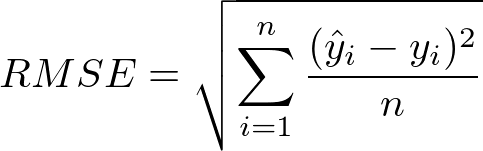
* The training data is first considered as vectors in multidimensional space (depending on features).
* There is a user-defined constant k, that’s why this algorithm is called K Nearest Neighbor or KNN.
* When a user passes an instance for classification, this algorithm first calculates the distance between the instance and all of the vector points given in the data. For this commonly euclidean distance is used, but other distance based methods can also be used, depending on the use case and performance.
* After calculating the distance between instance and all points in training data, the algorithm calculates its k nearest neighbors through sorting usually or some other method to find k nearest neighbors
* Then it performs majority voting usually on its k nearest neighbor, and classifies the given instance accordingly.

## Mean absolute error (MAE)



This is the most straightforward metric of evaluation known as Mean absolute error. The above is a fancy equation for evaluating it. It is literally the difference between what is an actual value to what our system predicts.

## Root mean square error (RMSE)

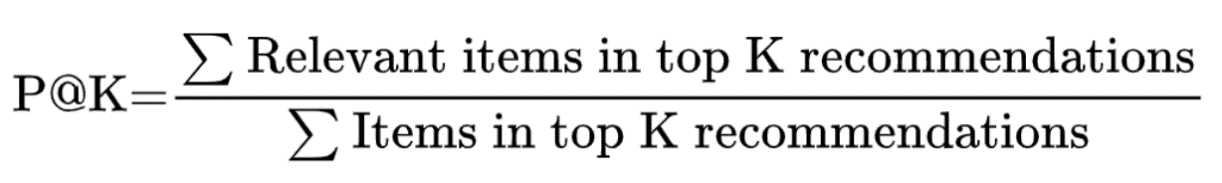


This is another common and perhaps most popular metric of evaluation.

One reason is that it penalizes you way less when you're close to actual prediction and way more when far from actual prediction compared to MAE.

**Precision @K**

Precision@K gives a measure of “out of K” items recommended to a user and how many are relevant, where K is the number of recommendations generated for a user.



## 3. Technical Review

**Pandas**

Pandas has within itself many rich data structures and features built to work with structured data is fast, easy, and expressive. The pandas have methods to describe manipulate, import, analyze, and even visualize the data. As you can see, it’s one of the important components that generally make Python a powerful and high-performing data mining environment in a self-sufficient way. The main object in pandas used for the most part, is a DataFrame, a two-dimensional, column-oriented, tabular data structure with row and column labels in a subtle way. Pandas provide really rich data structures and features designed to work with: Structured data is fast, easy, and expressive in a kind of big way. As you can see, it's one of the important components that make Python a powerful and high-performing data mining environment, or so they specifically thought. The main object in pandas used throughout this book kind of is a DataFrame, which is a two-dimensional, column- oriented, tabular data structure having rows and columns, similar to a Microsoft excel sheet.

**Matplotlib**

Matplotlib is the most widely used and most popular library in python for creating graphs and other 2D data visualizations. It is perfect for creating such stories that are well suited for publication. It integrates pretty well with IPython to provide a much user-friendly, interactive environment for building visualizations and performing exploratory data analysis.

**Seaborn**

Seaborn is a Python package for generating statistical graphs. It has a high-level interface with matplotlib and is tightly integrated with pandas’ inbuilt data structures.

The Seaborn library functions provide a declarative data- oriented API that converts data requests into graphs of their responses. Seaborn mechanically maps information values with visual characteristics such as color, size or style after running datasets and plotting specifications, generating internal transformation statistics and plotting decorations with useful axis labels and annotations. Many sea functions can generate numbers by comparing conditional subsets of information or concatenating completely separate variables in a data set. The purpose of seaborn is to be useful throughout the life of a scientific enterprise. Seaborn supports rapid prototyping and exploratory data analysis by building complete graphs from a single function call with the fewest parameters.

**Scikit-learn**

Scikit-learn is a Python module that integrates a great range of modern machine learning algorithms, and methods for supervised and unsupervised medium-scale problems. This package focuses primarily on bringing machine learning to non-experts using high-level general-purpose languages. Particular attention is paid to its easy nature of use, its high performance, rich documentation, and high API consistency. It has minimal to zero dependencies and is distributed under a simplified BSD license, making it suitable for both academic and commercial purposes

## 4. Requirements Specification

* 1. Hardware Requirement:
     + 500 GB hard drive (Minimum requirement)
     + 8 GB RAM (Minimum requirement)
     + PC x64-bit CPU
  2. Software Requirement:
     + Windows/Mac/Linux
     + Python-3.9.1
     + VS Code/ google colab/ Jupyter notebooks
     + Libraries:
       - Numpy 1.18.2
       - Pandas 1.2.1
       - Matplotlib 3.3.3
       - Scikit-learn 0.24.1

## 5. Conclusion:

* In this project we have built a system for recommending a product to user.
* For Model building we got the data from Shyft.
* We cleaned the data. Performed Label Encoding on the continuous variable
* Built a model on the cleaned dataset. Found out user based to be the best model.

## 6. Future Scope

* The project currently doesn’t have a content based filtering algorithm to recommend products, which can be implemented in future.
* This recommendation system is recommending clothing products currently. Using different data a similar system can be created for other products.
* Insights can be drawn to develop new products in future.

## 7. References

* + **ANALYSIS OF CLICKSTREAM DATA (2022)**

Animesh Jain, Ashish Kumawat

* + **Book: Machine learning using Python**

Manranjan Pradhan, U Dinesk Kumar