Project Report

of

User Behavior and Recommendation Analysis

Submitted by:

Name: Divyanshi Maurya

Reg. no.: 12219959

Section: K22UG

Roll no. : RK22UGA33

Submitted to:

Mr. Ved Prakash Chaubey (UpGrad, Instructor)

63892

Subject: CSE 353: EDA PROJECT

In partial fulfillment for the requirements of the award of the degree of

“Bachelors of Technology in Computer Science and Engineering”

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

TABLE OF CONTENTS

Project Title:

User Behavior and Book Recommendation Analysis Using Apriori Algorithm on Goodreads Data

1. Introduction:

In this report, we explore and analyze a collection of datasets derived from a **real-world e-commerce website.** The primary objective is to gain insights into user behavior and item properties through a thorough examination of the data. The datasets consist of four distinct files: e**vents_df, item_properties1_df, item_properties2_df, and category_tree_df.** This analysis aims to clean and manipulate the data, perform **exploratory data analysis (EDA)**, and draw meaningful conclusions to support subsequent decision-making and model-building processes.

2. Data Description:

The dataset **events_df** contains records of user interactions with items on the e-commerce platform. It includes the following columns:

- **timestamp**: An integer representing the time of the event.
- **visitorid**: An integer identifying the user involved in the event.
- **event**: A categorical variable describing the type of interaction (e.g., view, purchase).
- **itemid**: An integer representing the unique identifier of the item.
- **z_score_visitorid**: A float that has been computed to indicate the Z-score of visitorid.

The datasets item_properties1_df and item_propertie2_df describe properties associated with items, including:

- **timestamp**: An integer representing the time of the record.
- **itemid**: An integer identifying the item.
- **property**: A categorical variable describing the type of property (e.g., color, size).
- **value**: The value associated with the property (e.g., red, medium).

Lastly, category_tree_df provides hierarchical information about categories:

- **categoryid**: An integer representing the category.
- **parentid**: A float indicating the parent category ID, where NaN represents root categories.

3. Data Cleaning and Manipulation:

- The first step in data cleaning is to handle missing values and outliers. For the events_df, the transactionid column has a significant proportion of missing values (99.17%). This high percentage suggests that transactionid may not be crucial for the current analysis, so it can be safely dropped.

The **item_properties1_df** and **item_properties2_df** datasets do not have any missing values, which is ideal. However, further checks should be performed to ensure that the value fields contain consistent and valid entries.

The **category_tree_df** dataset has some missing values in the parentid column. These missing values are addressed by filling them with a placeholder value, such as -1, indicating a root category.

For outlier detection,

Z-scores are computed to identify extreme values. The Z-score is calculated using the formula:

$$Z=\sigma X-\mu \ / \ \mu$$

where X is the value, μ is the mean of the column, and σ is the standard deviation of the column. Values with Z-scores beyond a threshold (e.g., |3|) are considered outliers. For example, in the events_df, the Z-score of visitorid is used to identify outliers that may impact analysis.

Data normalization and encoding are then applied. Numeric columns, such as visitorid, are standardized using StandardScaler to bring them to a common scale. Categorical columns, such as event, are one-hot encoded to convert them into numerical format suitable for modeling.

Exploratory Data Analysis (EDA)

EDA begins with visualizing the data distributions. Scatter plots are employed to examine the relationship between visitorid and itemid. This visualization helps identify any patterns or anomalies in user interactions with items.

Pair plots provide an overview of the relationships between numerical variables, revealing correlations and trends. In the events_df, pair plots of visitorid, itemid, and z_score_visitorid highlight how these variables interact with each other.

Line charts are used to explore temporal trends. Plotting visitorid over timestamp helps detect any trends or periodic patterns in user activity over time.

Histograms are generated to understand the distribution of numerical variables. For example, histograms of visitorid, itemid, and z_score_visitorid show their frequency distributions, which can reveal skewness or other statistical properties.

The correlation matrix is computed to analyze the relationships between numerical features. A heatmap of this matrix displays how strongly different variables are correlated, assisting in feature selection and model building.

Additionally, count plots for categorical variables in item_properties1_df and item_properties2_df provide insights into the distribution of different properties and values.

Conclusion

The analysis reveals valuable insights into the e-commerce dataset. Key findings include the identification of outliers using Z-scores, which helps refine the dataset for more accurate modeling. The visualization techniques applied—scatter plots, pair plots, line charts, histograms, and correlation matrices—offer a comprehensive understanding of data distributions, relationships, and trends.

The cleaned and manipulated data is now ready for more advanced analyses or predictive modeling. By addressing missing values, handling outliers, and standardizing data, the dataset is prepared for robust analysis that can support informed decision-making and strategic planning.

Overall, the initial steps of data cleaning and EDA have laid a solid foundation for further exploration and modeling, ensuring that subsequent analyses are based on high-quality, well-understood data.