# CECS 550: Group 4 - Report Report: Analysis of E-commerce User Engagement for Item\_Id 481 – 640

Joseph Chorbajian, Aniruddha Gawande, Abhishek Jajoo, Satyam Sharma, and Ishan Unnarkar

## Introduction:

This report analyzes user engagement data to provide in-depth insights into customer behavior and preferences for a specific set of items on an e-commerce website. The data includes user actions (such as clicks, purchases, and views) and demographic information (such as age and gender). The analysis will provide insights into user behavior and preferences to predict whether the user will buy a certain item. This information can be also used to create a prediction engine for personalized recommendations and help identify potential improvement areas in the website's user experience.

## Analysis:

The following sections summarize the key findings of the analysis:

### Engagement by Action Type:

The data shows that clicks are the most common user action for the selected set of items, followed by views and purchases. This suggests that users are primarily interested in exploring products and browsing through different options before making a purchase decision.

Chart

Description automatically generated

### Gender Distribution:

Females make up more than 90% of all engagements for the selected set of items. This indicates that the website may have a predominantly female user base, and it may be beneficial to cater to their specific needs and preferences.

Chart, bar chart

Description automatically generated

### Clicks by Gender:

The data reveals that females are more likely to click on the selected set of items than males, by a significant margin. This highlights the importance of providing engaging and visually appealing product displays that can capture the attention of female users.

Chart, bar chart, histogram

Description automatically generated

### Age Distribution:

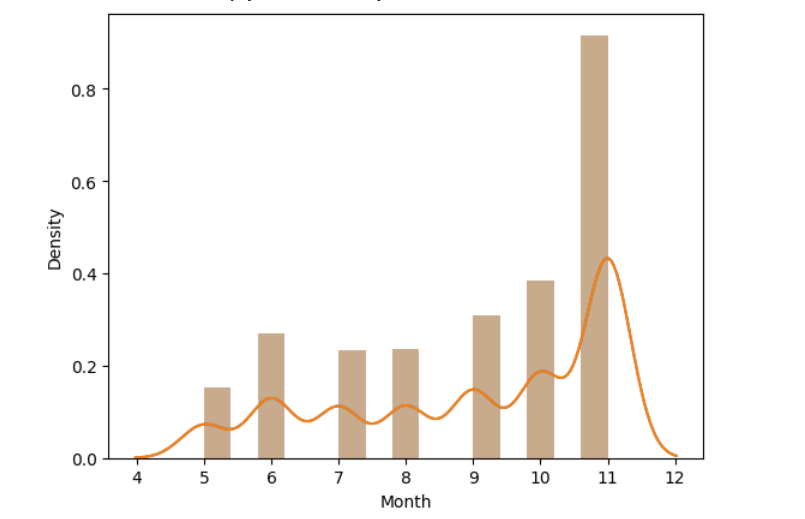
The data shows that users in age range 3 and 4 are the most active ones when it comes to engagement with the selected set of items, with age range 4 being the dominant one. This suggests that the website may have a specific target audience in terms of age group, and it may be useful to tailor the website's content and design to better suit their preferences.

Chart, histogram

Description automatically generated

### Active Months:

The data shows that November is the most active month of all, with the highest number of user engagements recorded during this month. This may be due to seasonal factors such as holiday shopping or other promotional events, and the website may need to consider similar events in other months to maintain user engagement throughout the year.



## New Features

New features were created for this dataset. Below is a summary of what was created:

|  |  |
| --- | --- |
| Feature | Description |
| Day | The extracted day from the time stamp. Ranges from 1 to 31. |
| Month | The extracted month from the time stamp. Ranges from 1 to 12. |
| Day of year | The day in the current year, taking into account months with different amounts of days. Possible ranges from 1-365, but actual ranges are from 162-345. |
| Day of week | The day of the week. Ranges from 0-6, where 0 is Sunday and 6 is Saturday. |

While binning the time and dates is important in getting an accurate model, additional features may be created from the data. Further analysis is required in order to determine what would help the models.

## Feature Rankings

The following models and algorithms were used to determine which features are considered the most important:

|  |  |
| --- | --- |
| Algorithm | Description |
| Random Forest | An ensemble classifier, fitting multiple decision trees in order to best explain the structure of the data. |
| Shapley analysis (on Random Forest) | Determines the importance of each feature, taking an average of its contribution to the output. |
| PCA | Reduces the number of dimensions on the data. While the output of PCA doesn’t correspond to a singular feature, it captures the most amount of features within the given number of components. |
| LDA | Linear discriminant analysis finds the most discriminating vectors for the data on a lower-dimensional dataset. The most important features will be those that separate the output the best. |

We note that a PCA model with 4 components can explain 27% of all the variance in the data while maintaining a similar accuracy to the other models, indicating that, while most other features vary, around 4 components are needed to describe it without any excessive loss.

For the other three models, we list the top features:

|  |  |  |
| --- | --- | --- |
| Random Forest | Shapley | LDA |
| Item ID | Age range | Gender |
| Day of week | Day of week | Action type |
| Category | Item ID | Month |
| Age range | Month | Day of week |
| Month | Category | Age range |

We see that month, age range, and day of week all appear in the top 5 features. Item ID and category also play a significant role.

## Prediction Model

We have currently tested the following models:

|  |  |
| --- | --- |
| Model | Test accuracy (20% split) |
| Random Forest | 92.3% |
| Random Forest (with one-hot encoding) | 92.4% |
| LDA | 92.7% |
| LDA (with some features dropped) | 93.1% |
| Naive Bayes Classifier | 93.2% |
| Naive Bayes Classifier (with one-hot encoding) | 92.6% |
| kNN (with k=6, Euclidean distance) | 93.3% |
| kNN (with k=6, Manhattan distance) | 93.1% |

We can see that kNN with 6 neighbors performs the best. Further tuning is required to get higher accuracies.

## Conclusion:

In conclusion, the analysis of the user engagement data for the selected set of items has provided valuable insights into user behavior and preferences on the e-commerce website. The findings suggest that the website may have a predominantly female user base, with a specific target audience in terms of age group. The website may benefit from providing engaging product displays, tailored to the preferences of these specific user groups. Additionally, seasonal factors such as holiday shopping and promotional events may be important in maintaining user engagement throughout the year.