Final Project Report (Three Pages)

Ruixin Gan
rg4743
rg4743@nyu.edu

Abstract

This project explores the user profile in online food delivery service based on Ele.me recommendation system. By clustering 1.8 million records of user behavior in an advertisement scenario, the project visualizes these group-level data with refined bubble chart, heatmap, and parallel set graph.

1. Introduction

In today's digital age, conducting complicated analytics is key to e-commerce success. This research introduces a subset from a unique dataset from Ele.me, focusing on user traits within their recommendation system over 1 day. It aims to shift the focus from traditional advertising Click-Through Rate (CTR) analysis to the online food delivery industry. Inspired by "MulUBA: multi-level visual analytics of user behaviors for improving online shopping advertising" [3], I used K-Means algorithm to cluster data and then visualize their traits in group level. By refining traditional bubble charts, heatmap, and parallel set diagram, this study examines how user behavior and other traits intertwine. This study also seeks to identify challenges and opportunities in enhancing recommendation systems, offering a new perspective on spatiotemporal user data analysis.

2. Results and Demonstration

The final result is an incline dashboard in Jupyter Notebook. By using Plotly for the graphs, the dashboard inherently supports various interactive operations like zooming, panning, triggering data series on and off, and displaying pop-ups with data point information on hover. These interactions allow users to explore the data more deeply and customize the view to their specific interests. The dashboard is structured into three rows.

These two bubble and bar charts in first row show groups' user behavior. Bars show the proportional distribution of each group within the total. The scatter(bubble) plot overlays the bar chart, where bubble size represents the value of specfic user behavior (CTR and order transacted for last 30 days). Bubble size in the first chart shows averages of behaviors, in the second shows totals. From comparing the bubbles from left to right we can infer the conversion rate of each group. By integrating bar charts with scatter plots within the same view, it allows observers to compare the size and performance metrics of different groups simultaneously.

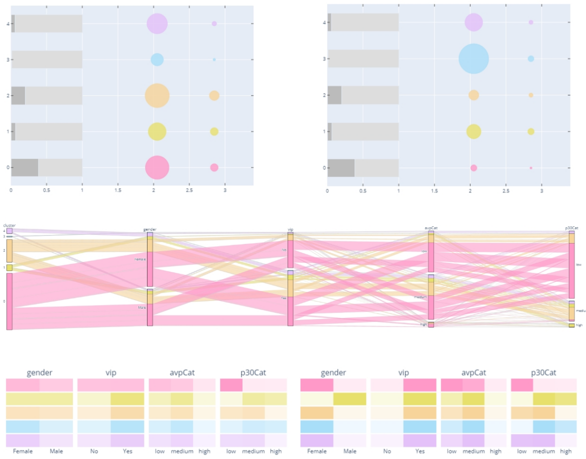

Figure 1. Dashboard

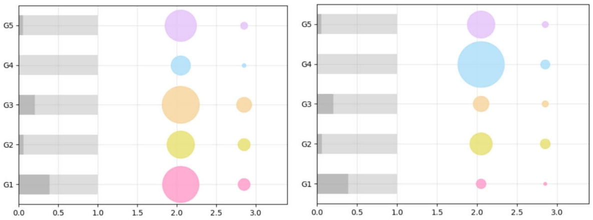

Figure 2. bubble and bar chart

Parallel set diagram, in the second row, shows how different groups relate to multiple user traits(gender, VIP status, spending level, and purchase frequency level), allowing the observers to trace and compare the pathways through these categories. Observers can dynamically adjust which dimensions they are viewing and how these dimensions interact [5]. Hovering over specific paths in the diagram highlights these pathways, providing additional insights into the distribution and frequency of data across different categorical combinations. For instance, group 0 is featured by having the highest price level but low purchase frequency users.

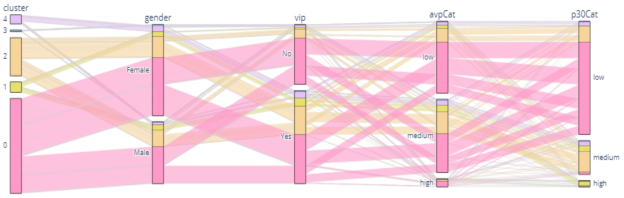

Figure 3. Parallel set diagram

Heatmap, in the last row, shows the group's proportion across these four user traits(gender, VIP status, historical average price and sum of price for last 30 days). Deeper colors mean more users in a category. The second heatmap introduces transparency into the color scale, making the lower values more visually subdued for comparison. Creating individual heatmaps for each group and arranging them in an orderly grid allows observers to compare patterns and trends across user trait catrgories horizontally.

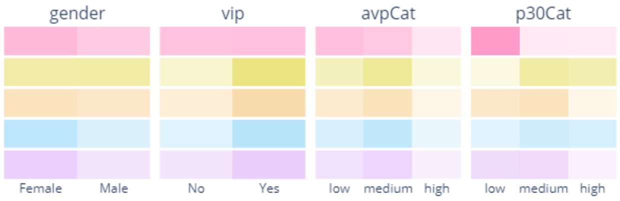

Figure 4. first heatmap

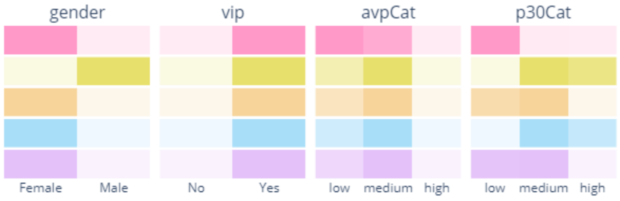

Figure 5. second heatmap

3. Implementation

The subset dataset used is from Ele.me dataset and include [1]:

a.User traits: gender, VIP status, historical average price, sum of price for last 30 days.

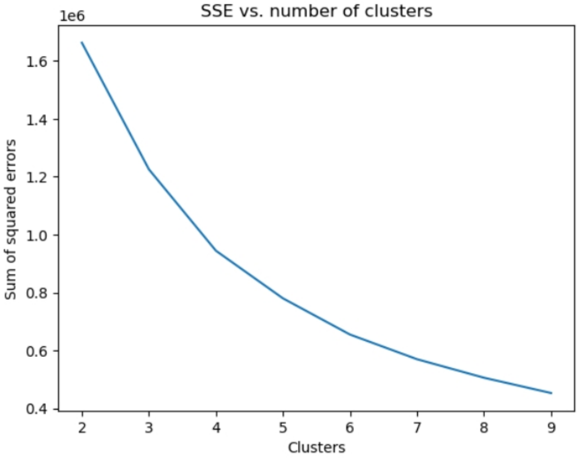

Figure 6. elbow method

b.User behavior: CTR indicators, activity (CTR and order transacted) for last 30 days

The first part is data cleaning. Select the first 1.8 million records from the subset dataset. Before starting visualization, redundancy and noise are eliminated. After standardization of 3 features of user behavior for cluster, the users are divided into 5 groups based on K-Means and the elbow method optimization. three rows.

The second part is visualization. The palette is set to gentle and diverse colors (light pink, pale yellow, light orange, light blue, and lavender). This is carefully chosen to ensure each group is distinctly recognizable, which suits quick understanding and comparison [6].

The first graph combines bar and scatter plots to present statistical data across different groups within the same axes. The size of the bubbles is adjusted using a scaling factor to fit the display scale appropriately, providing an intuitive sense of scale and magnitude [4].The scaling factor for bubble size and the offset constant are applied to position bubbles correctly after the bar components to the right side. Then the data categorization is done for all numeric variables of user traits with the mapclassify.NaturalBreaks classifier. The plotly.express.parallelcategories is applied to create the second graph, parallel categories diagram. This type of visualization is useful for showing multi-dimensional categorical data, revealing interactions between them.

The third graph, a complex heatmap, employs Plotly's make.subplots and Heatmap functionalities, specially designed for multi-category and multi-group data. Actually every row in a table is a single heatmap when plotting, but for visualizations it adopt a 5x4 grid layout, accommodating multiple heatmaps within a single figure. Minimal vertical/horizontal spacing (0.015) and the control for closing unnecessary x/y-axis labels optimize the appearance. Then a consistent color scale across the heatmaps particularly helps in making fair comparisons, as it keeps the visualization of data variations uniform across different groups.The different transparency design of two heatmaps are achieved by setting parameters zmin=0 and zmax=1 in the go.Heatmap function, furtherly helping in highlighting higher values more prominently, allowing for quicker visual identification of areas with more intense or significant metrics.

Finally, create an interactive dashboard using Dash, a popular Python framework for building web applications, integrated with Plotly for interactive visualizations [2].Two Matplotlib bubble and bar charts are converted into Plotly figures primarily. Bootstrap theme ensures a consistent and professional look in a responsive dashboard.

4. Discussion

By integrating multiple visualization toolkits, this project demonstrates how complex data can be transformed into a format that is easy to understand and manipulate. This multi-dimensional, interactive approach to visualization is highly effective for both internal data analysis and external presentation of insights.

For limitations, this study uses the environment of Jupyter Notebook, which makes the dashboard in an inline form, and some of the interactive graphics using Plotly results in a file size of up to 200M, which limit further optimization and adjustment of the dashboard layout. The original thesis (MulUBA) used Java script to create the web-side app, and the choice of interactions was flexible compared to Python.

For the extension of this study, the existing bubble bar chart and heat map should add interactive buttons to achieve the effect of switching between the same chart instead of displaying the two charts side by side, and it is recommended that Java script be used in combination with the front-end web design to increase this visualization.

References

[1] Recommendation data from ele.me. *Alibaba Cloud Tianchi*, 2022. 2

[2] Elias Dabbas. Interactive dashboardsand data apps withplotly and dash. 3

[3] Shangsong Liu, Di Peng, Haotian Zhu, Xiaolin Wen, Xinyi Zhang, Zhenghao Zhou, and Min Zhu. Muluba: multi-level visual analytics of user behaviors for improving online shopping advertising. *J. Vis.*, 24(6):1287–1301, dec 2021. 1

[4] Sackmone Sirisack and Anders Grimvall. Visual detection of change points andtrends using animated bubble charts. 2

[5] Zana Vosough, Marius Hogräfer, Loïc A. Royer, Rainer Groh, and Hans-Jörg Schulz. Parallel hierarchies: A visualization for cross-tabulating hierarchical categories. *Computers Graphics*, 76:1–17, 2018. 2

[6] Achim Zeileis, Kurt Hornik, and Paul Murrell. Escaping rgbland: Selecting colors for statistical graphics. *Computational Statistics Data Analysis*, 53(9):3259–3270, 2009. 2