Unveiling User Engagement:
A Data-Driven Analysis of Digikala

✉ Nima Azar
Student ID 99100285
nima.azar@sharif.edu

✉ Aref Namayandeh
Student ID 400108966
arefnemayandeh@gmai.com

✉ Hananeh Moballeghtohid
Student ID 400108911
hananehmoballeghtohid@gmail.com

Introduction: In the ever-evolving realm of e-commerce, user engagement stands as a fundamental pillar driving the success and sustainability of online platforms. Defined by a numerous of interactions and behaviors exhibited by users, user engagement serves as a key indicator of customer satisfaction, loyalty, and overall platform performance.

Recognizing the significance of user engagement in the digital landscape, businesses strive to define and measure meaningful metrics that encapsulate user interactions and sentiments within their platforms. Among these metrics, factors such as review length, the number of likes/dislikes each review received, the number of advantages/disadvantages mentioned, and other user-generated content parameters emerge as critical indicators of user engagement, providing valuable insights into user preferences, sentiments, and behaviors.

In this study, we embark on a comprehensive analysis to define and quantify metrics that hold significance to user engagement within an e-commerce platform known as Digikala. Leveraging rich datasets comprising product information, user reviews, and other pertinent variables, we aim to establish a comprehensive understanding of user engagement dynamics within the platform ecosystem.

Central to our analysis is the introduction of a novel parameter that represents user engagement for each product, derived from a combination of key engagement metrics such as review length, likes, and other relevant factors. This parameter serves as a measure of user engagement, enabling us to assess and compare the level of engagement across different products within the platform.

Furthermore, we endeavor to analyze this user engagement parameter in conjunction with various platform attributes such as product category, price, seller, brand, and more. Through statistical analysis and interpretation, we seek to unravel the relationship between user engagement and these variables, identifying patterns, trends, and actionable insights that can inform strategic decision-making and platform optimization efforts.

By delving into the interplay between user engagement metrics and platform attributes, we aim to provide valuable insights and recommendations for enhancing user experiences, optimizing product offerings, and fostering deeper connections with customers. Through our analysis, we endeavor to empower e-commerce platforms to drive sustainable growth and success in an increasingly competitive digital landscape.

Team Communication and Task Management

In our three-person team, effective communication and task management are vital for project success. We use various online platforms for seamless collaboration:

1. **Trello Workspace** We rely on Trello.com for task and time management. Its intuitive interface helps us organize tasks, assign responsibilities, and track progress using features like labels and due dates.

2. **GitHub for Code and File Sharing** GitHub serves as our platform for version control and code sharing. We use branches, pull requests, and code reviews to ensure code quality and consistency.

3. **Online Collaboration Sessions and Regular Feedback** Despite our small team size, we hold weekly online sessions to discuss project updates, share insights, and define tasks. Regular feedback ensures alignment with project objectives.

4. **Overleaf for Collaborative Writing** For collaborative writing and document preparation, we use Overleaf.com. Its real-time editing feature enables simultaneous contribution and review, streamlining the writing process.

Challenges

Our project faces several challenges that may impact its progress and outcomes:

1. **Low T-Ratios in Interpreting User Engagement Parameter:** When the parameter for user engagement shows a low t-ratio, likely due to high variance, it indicates that its coefficient estimate lacks statistical significance. This suggests that the parameter may not meaningfully contribute to explaining the variability in the data, making its interpretation unreliable for predicting the outcome variable.

2. **Data Cleaning Complexity:** Cleaning large datasets is time-consuming and complex, involving identifying and rectifying errors, missing values, and outliers.

3. **Complexity of Regression Explanation:** Explaining the intricacies of regression analysis and its implications to stakeholders with varying levels of statistical knowledge may present a significant challenge. Simplifying complex statistical concepts while retaining accuracy and relevance is essential for effective communication and understanding of our research findings.
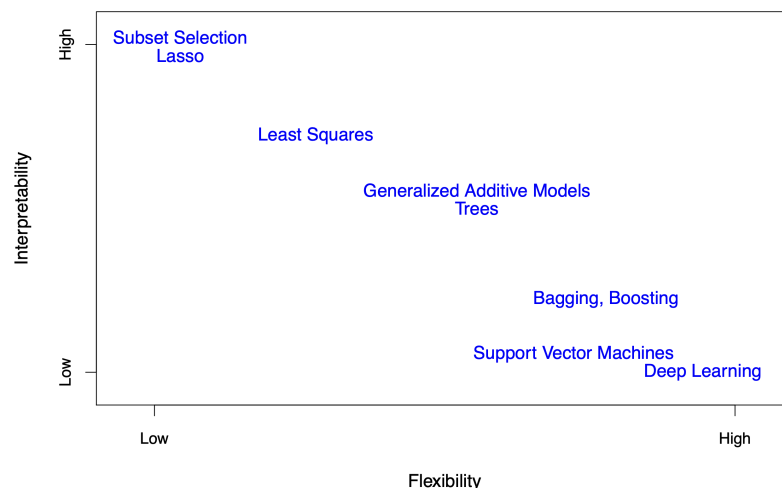

Figure 1. A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases

4. **Handling Large Datasets:** Processing and analyzing large volumes of data require robust computational infrastructure and efficient data management techniques.

5. **Project Quality and Time Constraints:** Ensuring the accuracy and rigor of our analysis while meeting project deadlines may require careful prioritization, resource allocation, and effective time management strategies.

Data Collection

The datasets utilized in this study were sourced from the Rade AI team and were originally published on LinkedIn. The dataset comprises comprehensive information on over 1.2 million products, including attributes such as price, product rating, number of votes, categories, brand names, and more. Additionally, it includes a vast collection of over 6 million product comments, featuring comment text, comment date, user ratings, and recorded advantages and disadvantages submitted by users.

The datasets are publicly available for download on both Kaggle and Hugging Face platforms, making them accessible to researchers and practitioners in the data field. The Rade AI"team collected this dataset for their business purposes in the summer of 1402 (2023) and subsequently released it for broader use within the data community.