# RECOMMENDATION SYSTEM ANALYSIS AND MODELING DOCUMENTATION

## GOAL

Personalize item suggestions for users based on their interaction history (clicks, add to cart, and transactions)

## A. ANALYSIS QUESTIONS

1. What common item properties are most frequently associated with items users add to cart after viewing?
   *(This will help understand what item features drive conversions from "view" to "addtocart".)*
2. How does the time spent viewing an item influence the likelihood of it being added to the cart or purchased?
   *(Calculate time gaps between events to explore this.)*
3. Are there users whose behavior patterns (e.g., excessively high number of clicks, zero add-to-cart or purchase) deviate significantly from typical user behavior?
   *(This will help define behavioral outliers.)*
4. Do abnormal users tend to interact with a specific subset of items or categories?
   *(Understanding this may help isolate bot traffic or fraud attempts.)*
5. What is the conversion funnel across the platform: from view → add to cart → transaction, and how does it vary across item categories or user types?
   *(This will help quantify behavior and optimize recommendation flow.)*
6. When do users engage most with the platform, and does this affect conversions?
   *(Analyze activity patterns across days of the week and hours of the day for views, add-to-cart, and transactions.)*
7. What are the most effective features for distinguishing normal from abnormal users?
   *(This will inform the feature engineering and model design for anomaly detection.)*
8. Can user viewing patterns be used to accurately predict the category of the item they are likely to add to cart?
   *(This addresses the core prediction problem using implicit signals.)*

## B. DATA PREPARATION

Before Preparation

- 3 Datasets;
  - Item_Properties.csv with 16M+ rows and 4 columns
  - Events.csv with 2.7M+ rows and 5 columns
  - Category_Tree.csv with 1,669 rows and 2 columns

- The data in the values column of the item_properties dataset had most of its content hashed, hence not representing actual values.

Cleaning

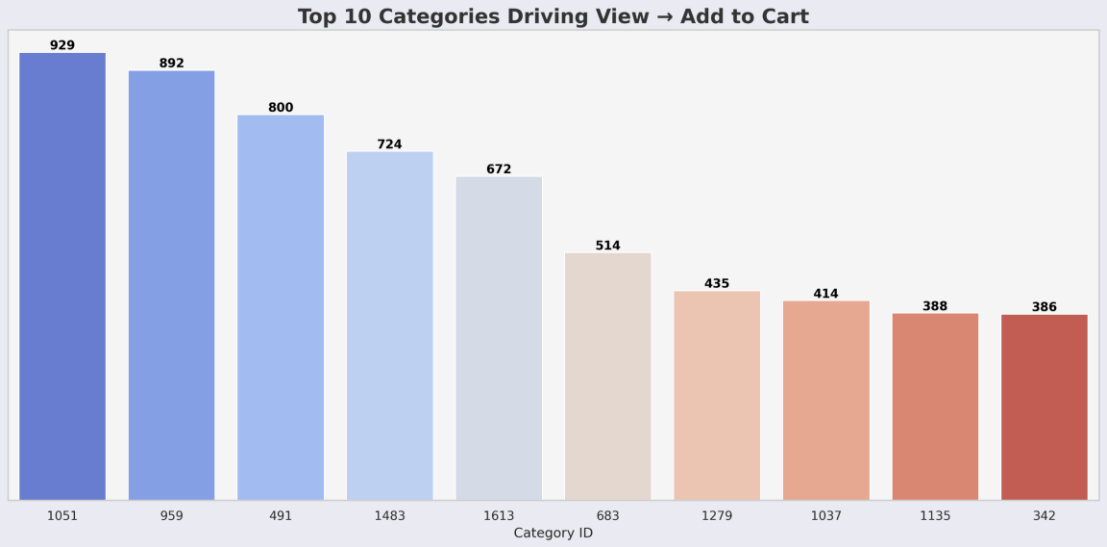
1. In the item_properties.csv dataset I filtered and kept only rows whose property field was "categoryid" or "available", and saved it as a .csv file. This was done because other rows apart from the ones kept, had their value field to be hashed values, which was inappropriate for analysis.
2. Converted timestamp column from object to datetime in all datasets.
3. Merged the events.csv dataset with the item_properties.csv dataset, pivoting on property fields, "categoryid" and "available", and saved it into a new .csv. file.
4. Finally merged the new .csv file with category_tree.csv dataset.

Dataset After Preparation

- About 2.8M rows, 8 parameters

## C. ANALYSIS, VISUALIZATIONS & INSIGHTS

1. What common item properties are most frequently associated with items users add to cart after viewing?



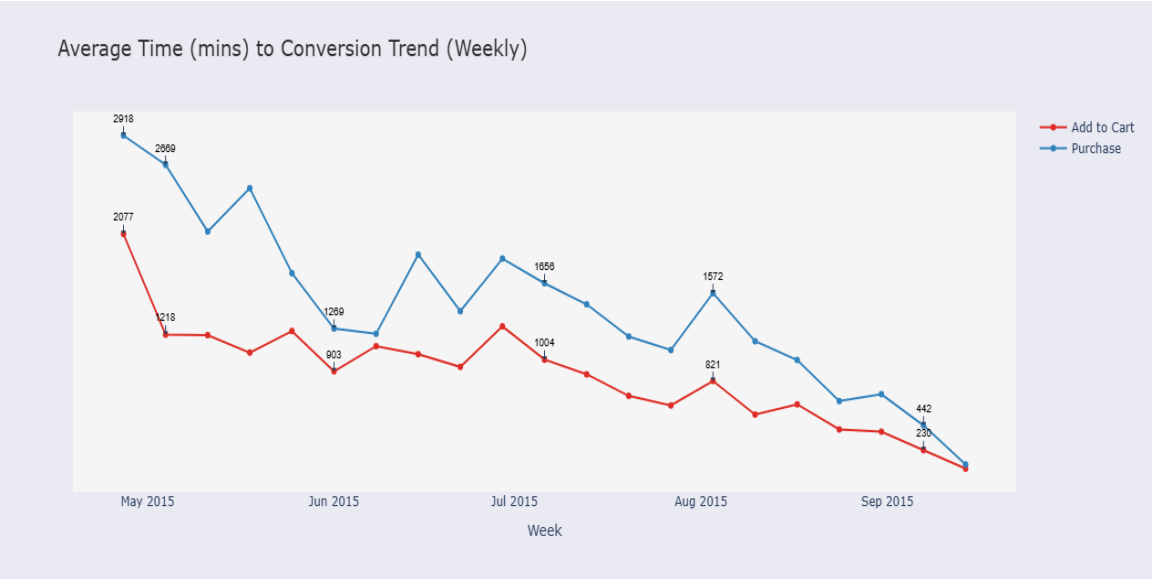

- The analysis of the top 10 product categories driving conversions from "view" to "add to cart" reveals clear engagement hotspots.
- Category IDs 1051, 959, and 491 emerge as the strongest performers, jointly accounting for a significant portion of add-to-cart actions.
- These categories appear to have higher product appeal or effective listing presentation, encouraging user action.

- The drop-off after the top five categories suggests opportunities for improving conversion rates in mid- and low-performing categories.

Key Takeaway

- High-performing categories can serve as benchmarks to uplift weaker segments through targeted strategies.

2. Can user viewing patterns be used to accurately predict the category or price range of the item they are likely to add to cart?



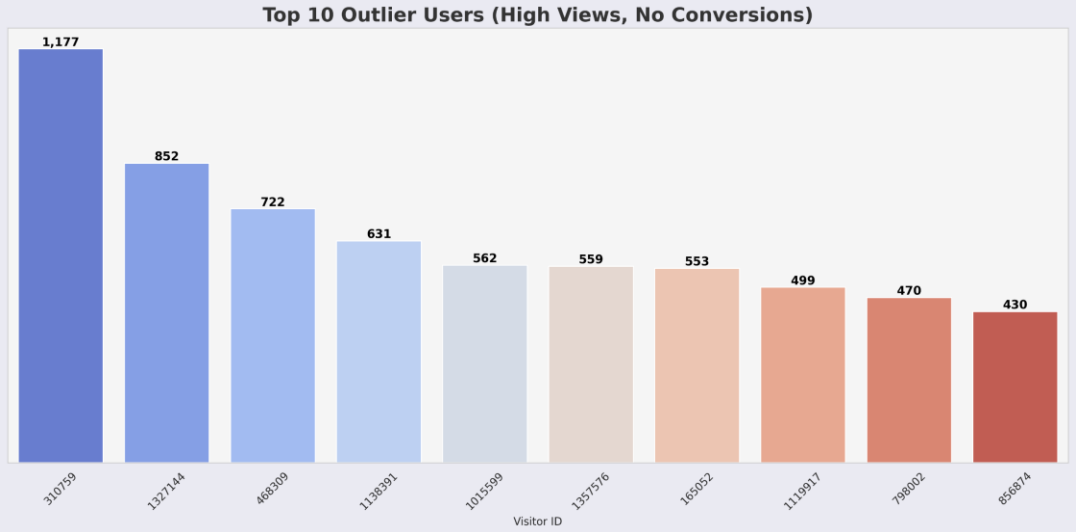

- The weekly trend shows a clear decline in the average time users take to convert from a product view to adding it to the cart or completing a purchase.
- At the start of the period, purchase conversions averaged over 2,900 minutes, while add-to-cart actions averaged above 2,000 minutes, suggesting slower decision-making.
- By the final weeks, both metrics dropped sharply, with purchases averaging under 500 minutes and add-to-cart actions around 230 minutes, reflecting quicker user decisions.
- This decline may be linked to improved product recommendations, targeted promotions, or better user familiarity with the platform.

Key Takeaway

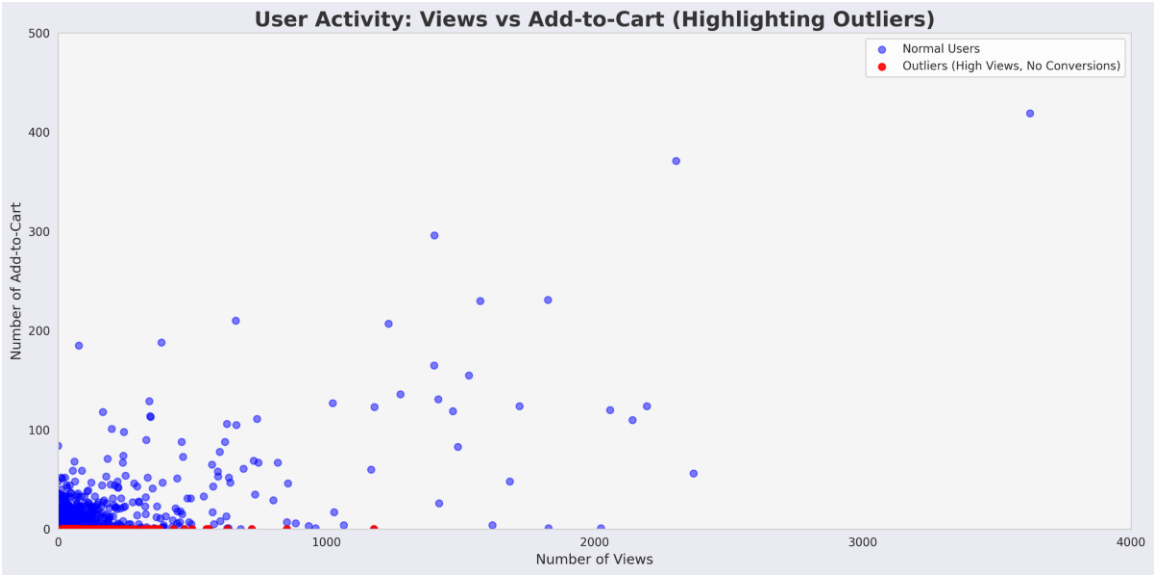
- Conversion times have steadily decreased across the period, with purchase delays dropping by over 80%.
- Faster add-to-cart and purchase actions indicate growing platform efficiency and user decisiveness.

- Users are making faster buying decisions, signaling improved engagement and possible gains in recommendation quality.

3. Are there users whose behavior patterns deviate significantly from typical user behavior?




- The chart highlights the top 10 visitor IDs with unusually high product view counts but zero conversions (no add-to-cart or transactions).
- The leading outlier, Visitor ID 310759, recorded an extreme 1,177 views, far exceeding the next highest at 852 views.
- This pattern suggests possible browsing-only behavior, bot-like activity, or intent-driven searches that failed to convert.
- Identifying these users is critical for targeted re-engagement strategies, bot detection, and improving conversion funnels.

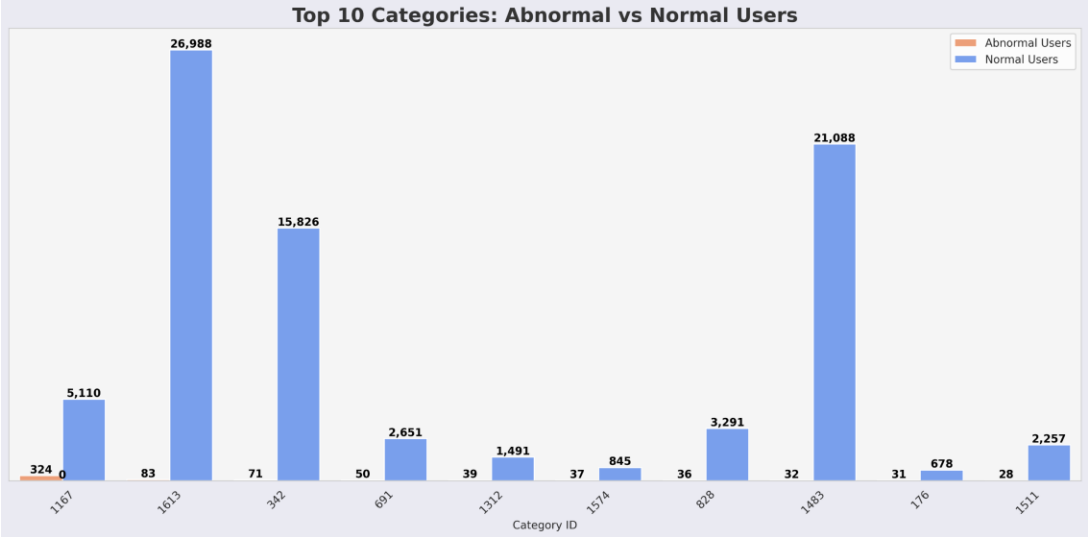


- Most users cluster in the bottom-left quadrant with fewer than 500 views and 100 add-to-cart actions, indicating typical browsing-to-cart behavior.
- Outliers are characterized by extremely high views (up to ~3,500) but zero add-to-cart activity, strongly suggesting disengagement or possible automated browsing.
- The lack of add-to-cart actions from these high-view users skews platform engagement metrics and may signal opportunities for bot filtering or targeted re-engagement.

Key Takeaway

- A small group of visitors accounts for disproportionately high views without converting, indicating potential user disengagement or non-genuine activity that needs targeted action.
- Outlier detection reveals high-engagement but zero-conversion users, crucial for refining recommendation and fraud detection systems.

4. Do abnormal users tend to interact with a specific subset of items or categories?




--- Category Share Comparison ---

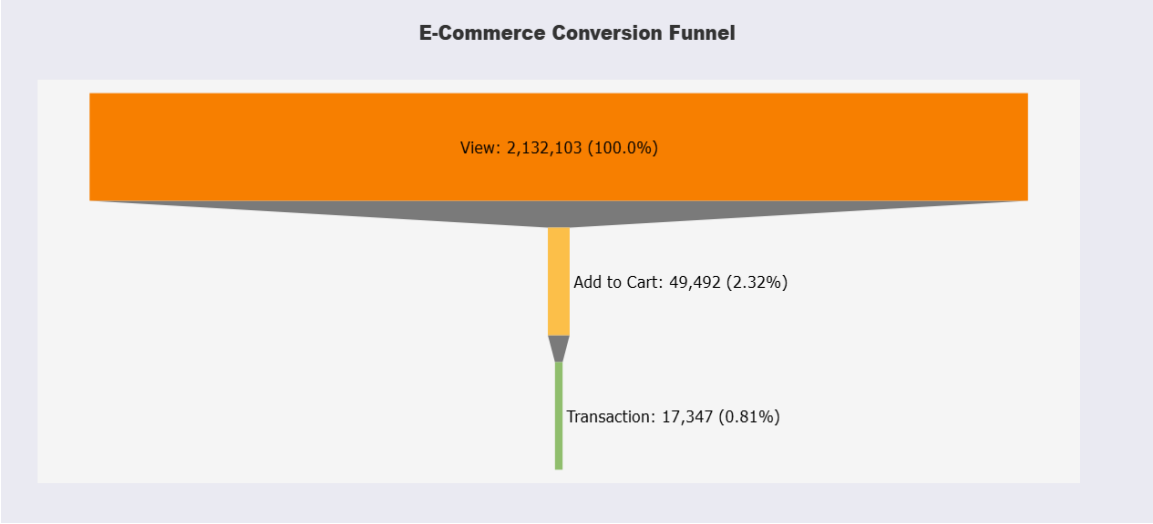
| user_type | categoryid | Abnormal Users | Normal Users | Abnormal % | Normal % |
|---|---|---|---|---|---|
| 0 | 1167 | 324 | 5110 | 44.32 | 6.37 |
| 5 | 1613 | 83 | 26988 | 11.35 | 33.64 |
| 7 | 342 | 71 | 15826 | 9.71 | 19.73 |
| 8 | 691 | 50 | 2651 | 6.84 | 3.30 |
| 1 | 1312 | 39 | 1491 | 5.34 | 1.86 |
| 4 | 1574 | 37 | 845 | 5.06 | 1.05 |
| 9 | 828 | 36 | 3291 | 4.92 | 4.10 |
| 2 | 1483 | 32 | 21088 | 4.38 | 26.29 |
| 6 | 176 | 31 | 678 | 4.24 | 0.85 |
| 3 | 1511 | 28 | 2257 | 3.83 | 2.81 |

- The category distribution comparison reveals stark behavioral differences between abnormal and normal users.
- Abnormal users are heavily concentrated in category 1167 (44.3% of their activity), which represents only 6.4% of normal user activity, suggesting a targeted or automated browsing pattern.
- Conversely, categories like 1613 (33.6%) and 1483 (26.3%) dominate normal user interactions but account for just 11.4% and 4.4% respectively among abnormal users.
- This skewed engagement profile indicates that abnormal users disproportionately interact with a narrow set of categories, likely reflecting non-typical browsing behavior that may be linked to bots or fraudulent activity.

Key Takeaway

- Abnormal users show a highly skewed focus on a few categories, especially 1167, which forms nearly half their activity but is minor for normal users.
- Normal users engage more evenly, with dominant interest in categories 1613 and 1483, showing a broader browsing pattern.
- Abnormal users' category engagement is narrow and atypical, signaling potential automated or fraudulent activity.

5. What is the conversion funnel across the platform: from view → add to cart → transaction, and how does it vary across item categories or user types?



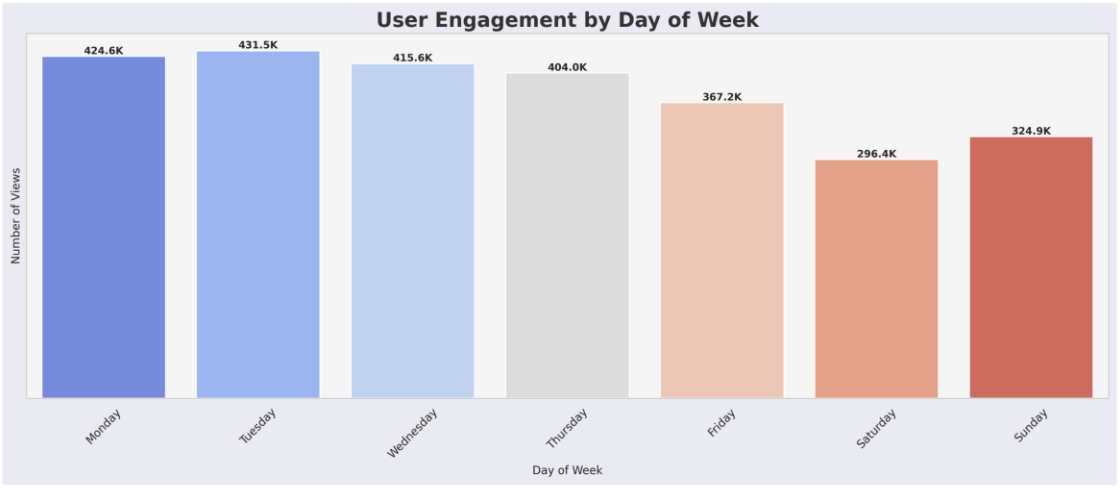

- The funnel reveals a significant drop-off from product views to subsequent actions.
- Out of 2,132,103 unique item views, only 49,492 (2.32%) progressed to the add-to-cart stage, indicating low initial engagement beyond browsing.

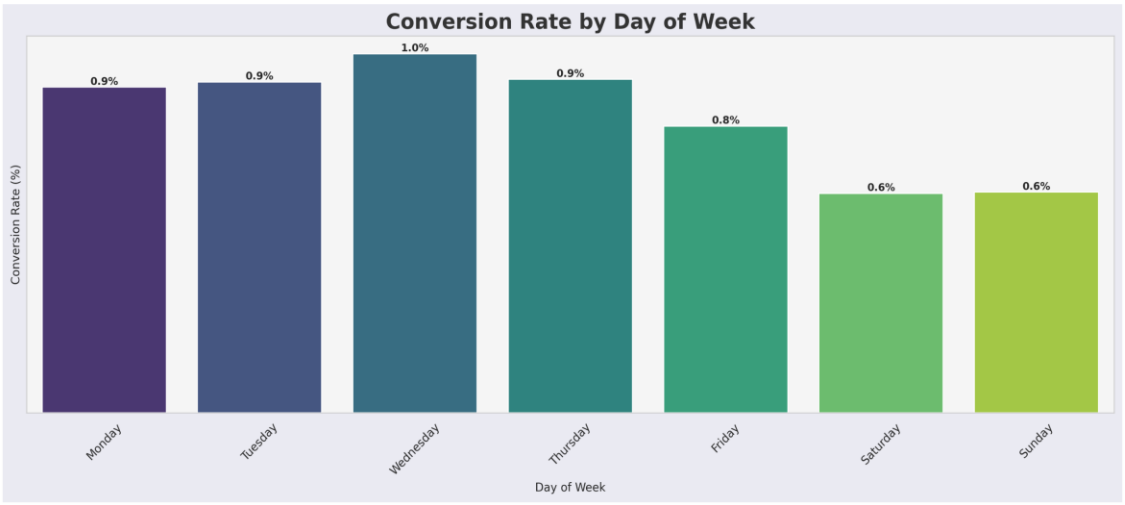
- The final conversion to purchases is even lower, with 17,347 transactions, representing just 0.81% of all views.
- This steep attrition suggests potential issues in product appeal, pricing, or checkout experience.

Key Takeaway

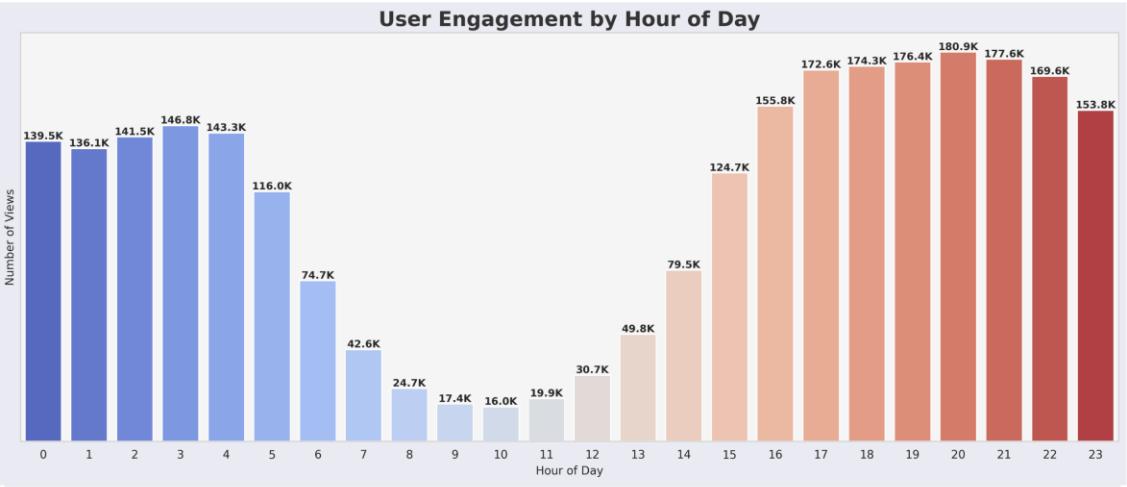
- The platform loses over 97% of potential customers before they even add items to the cart.

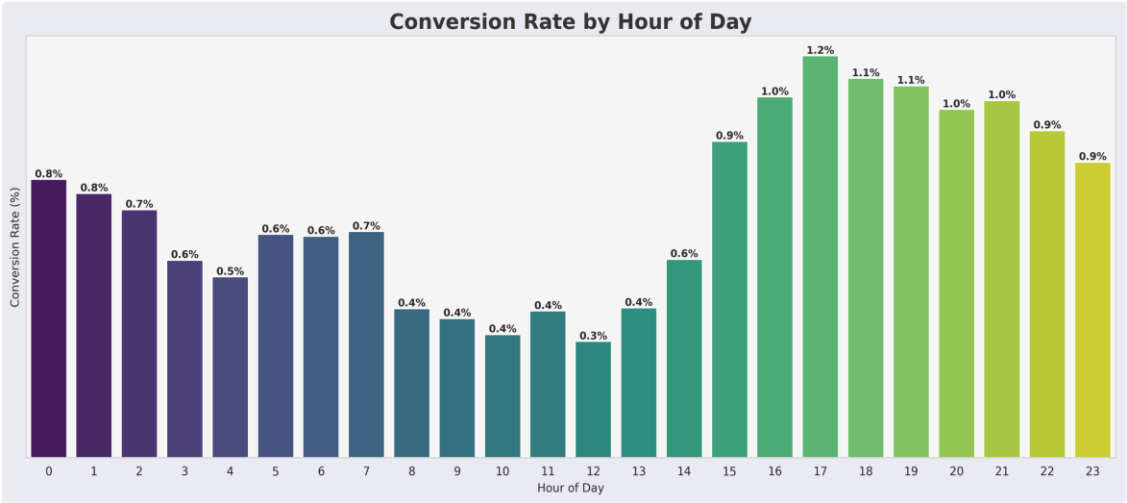
6. When do users engage most with the platform, and does this affect conversions?




- User engagement is highest on Mondays, Wednesdays, and Thursdays, with Monday leading slightly in total views.
- Engagement drops toward the weekend, particularly on Saturdays and Sundays.
- Conversion rates, however, peak midweek on Wednesday (9.1%) and Thursday (8.9%), indicating these days are most effective for driving purchases.
- This suggests a mismatch between engagement and conversion on certain days, with midweek offering the best balance of traffic and transaction efficiency.



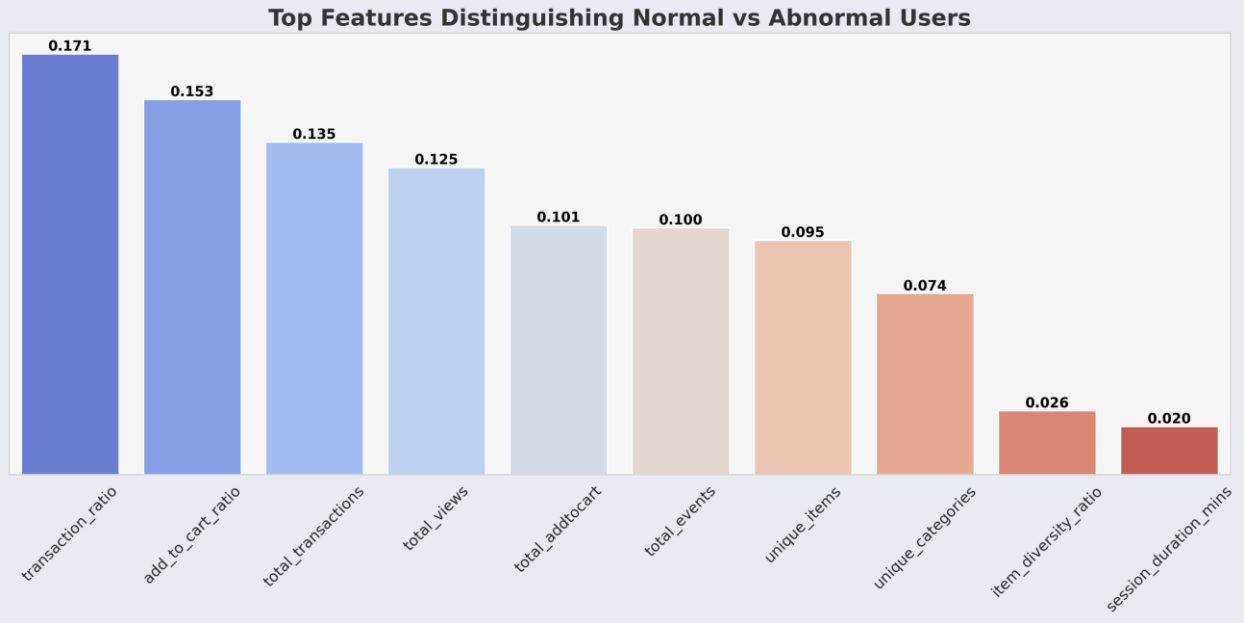




- User engagement peaks between 18:00 and 21:00, with the highest activity around 20:00, while the lowest activity is between 02:00 and 06:00.
- Conversion rates are highest from 19:00 to 23:00, peaking at 21:00, showing a strong correlation between late-evening activity and purchases.
- Weekends drive higher engagement than weekdays, particularly Saturday and Sunday, while Tuesday and Wednesday show lower activity.
- This pattern suggests optimal campaign timing in late evenings and weekends to maximize both reach and conversions.

Key Takeaways

- User interactions show clear temporal and behavioral patterns.
- Users decide faster as weeks progress, especially for add-to-cart actions.
- Activity and conversion rates peak midweek (Wednesday–Thursday), suggesting these days are optimal for running targeted sales or promotions.

7. What are the most effective features for distinguishing normal from abnormal users?




- The analysis reveals that transaction ratio is the strongest differentiator between normal and abnormal users, suggesting that purchase completion patterns are the clearest signal for detecting unusual behavior.
- This is followed by add-to-cart ratio and total transactions, indicating that abnormal users tend to have significantly different carting and buying habits compared to normal users.
- Engagement metrics like total views and total add-to-cart also contribute meaningfully, but less so than conversion-related ratios.
- Interestingly, session duration and item diversity ratio have minimal importance, implying that browsing length and product variety are weak indicators of abnormality.

Key Takeaway

- Focus on **conversion ratios** and **transaction counts** as primary features for anomaly detection, as they provide the clearest separation between user types.

8. Can user viewing patterns be used to accurately predict the category of the item they are likely to add to cart?

Random Forest Modeling was implemented, because it works best on nonlinear relationship data. It is able to predict more accurately, because it trains multiple decision trees. The results were as follows.

| Metric | Score | Interpretation |
|---|---|---|

| | | |
|---|---|---|
| Top-1 Accuracy | 0.8499 | The model predicts the correct category as its first choice in about 85% of cases, showing strong primary prediction accuracy. |
| Top-3 Accuracy | 0.9214 | In over 92% of cases, the correct category is among the model's top 3 predictions, ensuring highly relevant suggestions even if the first guess is wrong. |
| Macro Precision | 0.5076 | When the model predicts a category, it is correct roughly 51% of the time across all categories, reflecting balanced precision for both frequent and rare classes. |
| Macro Recall | 0.3920 | The model retrieves about 39% of actual categories correctly across all classes, indicating room for improvement in identifying all possible relevant categories. |
| Macro F1-score | 0.4244 | The balance between precision and recall across all categories is 42%, suggesting generally good but improvable consistency in predictions for both common and rare categories. |

- Strong Predictive Signal – User viewing patterns contain enough information to accurately predict the category of items they are likely to add to cart.
- Balanced Category Coverage – The precision (~51%) and recall (~39%) indicate the model is not overly biased towards only the most popular categories; it performs reasonably well even for less common ones.
- Room for Improvement in Recall – The recall score shows the model misses ~61% of actual categories across all classes, meaning some user intents are not being captured.
- Recommendation Potential – The high Top-3 accuracy suggests the model can be confidently used in a recommendation system where multiple category suggestions are displayed.

Recommendations Prediction App

The recommendation system predicts the category id of the item a user is likely to add to cart based on four user session behaviours as inputs. The model encodes these features and runs them through the trained Random Forest model. The output gives the predicted category + top 3 alternatives with probabilities.

**Input Fields**

- Last Viewed Category ID: This is the category of the most recent product the user viewed before potentially adding something to their cart. If the ID typed was never seen during training, the system automatically maps it to "Unknown_CatID".
- Most Frequently Viewed Category ID: This is the category the user viewed most often across their browsing session. Again, unseen IDs get replaced with "Unknown_CatID" safely.
- Number of Unique Categories Viewed: This measures how diverse the user's browsing behavior is. A higher number means the user is browsing broadly, while a lower number means they're focused.
- Total Views Before Add-to-Cart: This is the total number of product views a user made before deciding to add something to their cart. This helps the model capture how long a user takes to make a decision.

**Output Fields**

- Predicted Category: This is the single most likely category the system predicts the user will add to their cart.
- Top 3 Predictions (Category, Probability): This gives the 3 most likely categories along with probabilities, so the recommendation is not limited to one guess.

## D. GENERAL INSIGHTS & CONCLUSIONS

1. User behavior patterns carry strong predictive signals. Viewing patterns (recent views, frequency, and browsing diversity) are highly correlated with the items users eventually add to cart, enabling accurate category predictions (~85% Top-1 accuracy).
2. Conversion is highly selective. Out of millions of item views, fewer than 3% progress to add-to-cart, and <1% result in purchases — indicating major drop-offs in the funnel.
3. Temporal trends shape engagement. User activity and conversions peak midweek (Wednesday–Thursday) and evenings (19:00–23:00), which represent the most effective engagement windows.
4. Abnormal user behavior is distinct. Outliers often show very high view counts with zero conversions, and are concentrated in narrow item categories (e.g., Category 1167). This highlights the importance of anomaly detection for fraud prevention and platform health.
5. Category bias exists. Some categories (1051, 959, 491) drive disproportionate conversions. These can act as benchmarks for weaker categories.

## E. RECOMMENDATIONS

1. Deploy the Recommendation System: Integrate the trained Random Forest model into the platform to provide real-time category predictions, showing Top-3 suggestions to maximize relevance and conversion.
2. Focus on Conversion Optimization: Investigate reasons for the steep funnel drop-off (pricing, checkout friction, product detail clarity), and benchmark against top-performing categories.
3. Leverage Timing for Campaigns: Run promotions and recommendations in midweek evenings to align with peak activity and higher conversion rates.
4. Handle Abnormal Users Proactively: Build filters for bot-like behavior (excessive views, no conversions, skewed category focus) to improve data quality and reduce fraud.
5. Enrich datasets with metadata (category names, product attributes) for richer analysis & clearer insights.

Github link for full project: https://github.com/dspselorm/recommendation-system-analysis

Recommendation System App Link: https://huggingface.co/spaces/selorm-etse-forfoe/ecommerce-item-category-recommendation-system