Analyzing User Journeys in E-Commerce

Vipul Mishra, Abhishek Bansal

Indraprastha Institute of Information Technology, Delhi

Abstract

This report analyzes user conversion funnels in an e-commerce platform to understand how user behavior influences conversion rates. Using a dataset with over 2.7 million user events and 20 million item property records, we explore challenges in handling large data, apply dimensional- ity reduction techniques, and compare machine learning models on scaled and unscaled datasets. Hypothesis testing reveals insights into user interactions, while model performance comparisons demonstrate the benefits of scaling for improved efficiency. The findings can help optimize marketing strategies and enhance user experience, directly contributing to increased revenue.

1 Dataset Overview and Problem Statement

The dataset, available at
https://www.kaggle.com/datasets/retailrocket/ecommerce- dataset/data,
consists of two primary components:

- Events Data: Contains 2.7 million user interactions labeled as ‘view‘, ‘add to cart‘, or ‘transaction‘.
- Item Properties: Includes over 20 million records describing the characteristics of more than 400,000 unique items.

Problem Statement: The goal is to analyze user conversion funnels to understand how user behavior affects conversion rates from views to purchases.

2 Challenges and Solutions

2.1 HighDimensionality

One of the primary challenges was handling the high dimensionality of the dataset. With millions of data points and multiple features, training machine learning models became computationally expensive. To address this, we employed the Johnson-Lindenstrauss Lemma (JL Lemma), a mathematical technique for dimensionality reduction.

Johnson-Lindenstrauss Lemma: The JL Lemma allows for reducing the dimensionality of high-dimensional data while approximately preserving pairwise distances between data points. This technique is particularly useful for datasets with high-dimensional features like ours, which contains over 20 million records.

Benefits of JL Lemma:

- Faster Training :Byreducingthefeaturespace,thecomputationalandmemoryrequirements for model training decrease significantly.
- Minimal Accuracy Loss: The structure of the data is preserved, with accuracy dropping by only 2-3%.
- Efficiency Gains: Lower model complexity leads to faster predictions and more efficient use of computational resources.

2.2 Large Data Volume

The sheer volume of data presented another challenge. The events dataset contains 2.7 million interactions, while the item properties dataset has over 20 million records. Processing such large datasets required careful handling to avoid memory issues and ensure efficient computation.

Solutions Implemented:

- Handling Missing Values :Rowswithmissingvaluesincriticalfieldsweredroppedto maintain data integrity and reliability.
- Encoding Categorical Features: The ‘event‘ feature was converted into numerical representations (e.g., 0 for ‘view‘, 1 for ‘add to cart‘, and 2 for ‘transaction‘).
- Stratified Sampling: To create balanced datasets for training, we used stratified sampling, ensuring that each class (view, add to cart, transaction) was proportionally represented.

3 Hypothesis Tests and Results

3.1 Effect of Item Category on Add-to-Cart Rate

We conducted a hypothesis test to determine if certain item categories are more likely to be added to the cart than others.

Hypotheses:

- Null Hypothesis (H_0): The add-to-cart rates are evenly distributed across item categories.
- Alternative Hypothesis (H_1): There is a significant difference in add-to-cart rates based on item categories.

Test Performed: Chi-square test for goodness-of-fit.

Result: The null hypothesis (H0) was rejected, indicating significant differences in add-to-cart rates among different item categories. This suggests that certain categories are more appealing to users, influencing their likelihood to add items to the cart.

3.2 Conversion Rates: New vs. Returning Users

We also examined whether new users are as likely to convert (i.e., make a purchase) as returning users.

Hypotheses:

- Null Hypothesis (H_0): Conversion rates are the same for new and returning users.
- Alternative Hypothesis (H_1): Conversion rates differ significantly between new and returning users.

Test Performed: Z-test for two proportions.

Result: We failed to reject the null hypothesis (H0). This indicates that there is no statistically significant difference in conversion rates between new and returning users. Both user types exhibit similar purchasing behaviors.

4 Feature Extraction and Engineering

To predict if a user will purchase an item, we extracted features that influence user behavior and conversion:

- Visitor ID: Helps differentiate between new and returning users.
- Item ID: Identifies specific products.

- Day and Hour: Temporal features representing when the item was viewed.
- Item Properties: Includes changing attributes like price and category that affect purchase likelihood.

The target variable is whether the user completed the purchase funnel within a week of viewing the product.

5 ModelTrainingandEvaluation

Three machine learning models were used for prediction:

- Random Forest Classifier: 1000 trees, testing accuracy: 66.54%.
- Multi-Layer Perceptron (MLP): Two hidden layers (100 and 50), accuracy: 56.62%.
- Gradient Boosting Machine (GBM): 1000 estimators, accuracy: 58.98%.

5.1 Using Johnson-Lindenstrauss Lemma for Scaling

The JL Lemma reduced dimensionality while preserving data structure. This resulted in significant time reductions and minor accuracy loss (2-3%).

6 Comparison of Model Performance on Scaled vs. Unscaled Datasets

6.1 Training Time Analysis

Table 1: Training Time Comparison (in seconds)

Model	Unscaled Dataset	Scaled Dataset	Time Reduction (%)
Random Forest (RF)	617.84	431.48	30.2%
Gradient Boosting (GBM)	4124.59	2153.79	47.8%
Multi-Layer Perceptron (MLP)	3254.36	2076.51	36.2%

6.2 Accuracy Comparison

Table 2: Accuracy Comparison on Scaled vs. Unscaled Datasets

Model	Unscaled Accuracy (%)	Scaled Accuracy (%)
Random Forest (RF)	66.54	61.69
Gradient Boosting (GBM)	58.98	58.60
Multi-Layer Perceptron (MLP)	56.62	49.58

Observation: Training time reduced significantly, while accuracy dropped marginally.

7 Images and Visualizations

We include key visualizations to support our analysis:

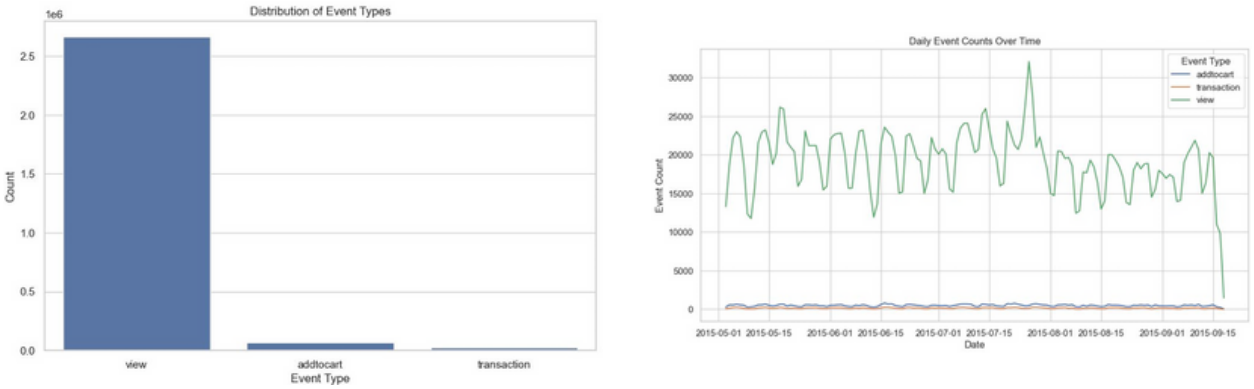

Figure 1: Funnel Analysis: User interactions across different stages (View, Add to Cart, Transaction).

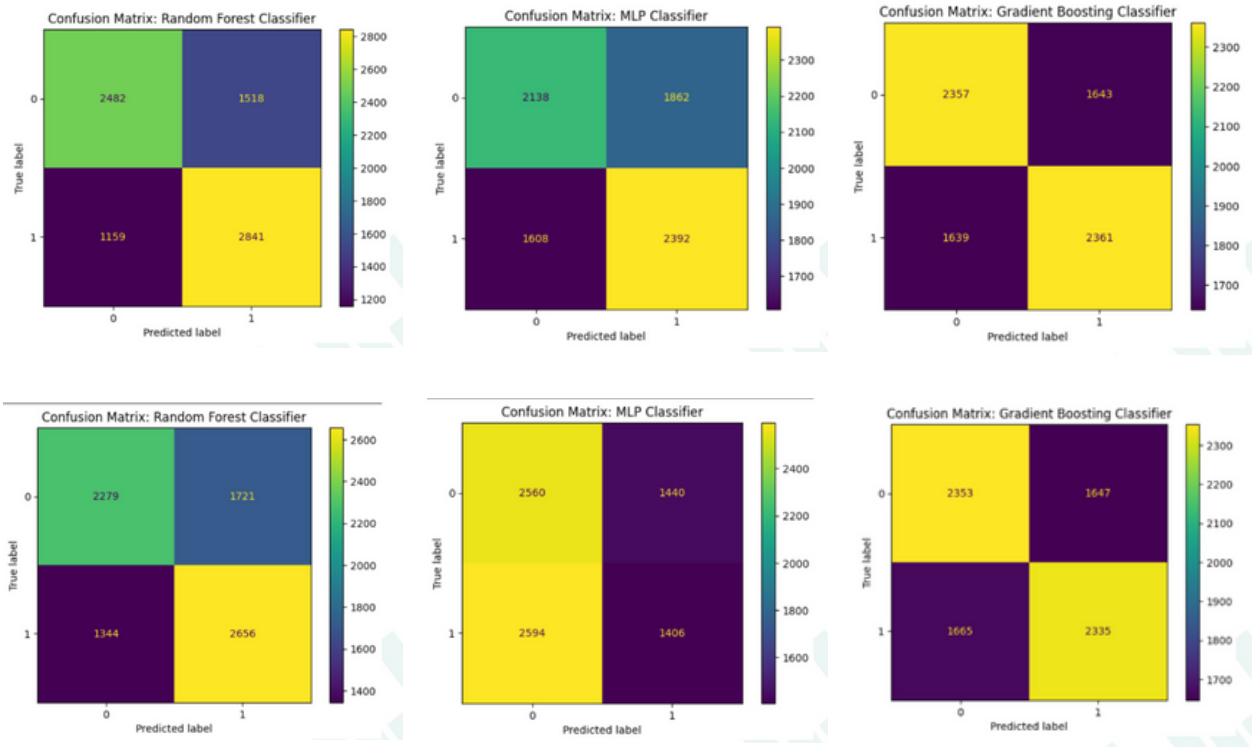

Figure 2: Results of ML model training.

8 Conclusion

By applying dimensionality reduction using the JL Lemma, we achieved faster training times with minimal loss in accuracy. Random Forest and Gradient Boosting performed best, balancing accuracy and efficiency.

Future Work:

- Analyze advanced deep learning models, such as CNNs or RNNs, to improve predictive accuracy.
- Investigate temporal and session-based patterns for personalized recommendations.