IAT461 - Project Proposal
Stephy Hin Kiu Wong (301452605)
Hinata Nozawa (301463370)

PROJECT DESCRIPTION

This project aims to build a classification model that predicts whether a user will make a purchase based on their interactions with a product on an e-commerce platform. Using a large-scale behavioural dataset containing over 2.7 million event records, the model analyzes user-product interaction patterns to estimate the likelihood of conversion. We will be engineering features that capture user engagement patterns, such as number of views and add-to-carts, the time between events, total interaction count, event sequences, etc. The target variable is binary, indicating whether the session resulted in a completed transaction. The target users for this project would be the e-commerce marketing team and data analysts, as it not only identifies key behavioural signals of purchase intent but also illustrates how data-driven insights can enhance recommendation systems and inform personalized marketing strategies in digital commerce environments.

RESEARCH QUESTION

Can we predict whether a user will make a purchase based on their interactions with an item? (Yes/No)

Target: event, event=’transaction’ returns 1, else, returns 0
Predictors: number of 'view' and 'addtocart' events per item/user, time between first view and last event, total number of interactions, event sequence patterns, etc.

DATASET

events.csv from the Retailrocket recommender system dataset

EARLY EXPLORATION SUMMARY

Context
Events.csv is a behaviour data that logs the timestamp, user ID, item ID, the type of event (view, add_to_cart, or transaction), and a transaction ID for purchases, with a total of 2756101 records.

Content

Column	Type	Description
timestamp	int64	The time of the event in Unix milliseconds
visitorid	int64	User identifier
event	object	The type of interaction Categorical: view /add_to_cart / transaction
itemid	int64	Product identifier
transactionid	float64	Only populated when a transaction is recorded

User and Item Diversity:

Unique visitors	1,407,580
Unique items	235,061

Event Breakdown:

view	2,664,312 (96.6%)
add_to_cart	69,332 (2.5%)
transaction	22,457 (0.8%)

(The skew shows that purchases are rare events, potentially making this an imbalanced classification problem.)

Session Behaviour

An initial look at the data, by grouping events by user and calendar day, suggests variation in how users interact with the platform. Many users appear to engage in brief sessions with a smaller number of events, but some users perform several actions in a day. This indicates the potential to investigate session length and intensity as behavioural signals. The analysis could explore whether longer or more frequent sessions are correlated with purchase behaviour, or whether session characteristics can help distinguish between casual browsers and high-intent buyers. Session-based features, such as average session length or peak interaction times, may prove valuable in building classification models.

Temporal Dynamics

Temporal patterns in user behaviour also present a potential for analysis. By examining the time gaps between consecutive events, we can begin to distinguish between users who act quickly and those who return after long intervals. Early observations suggest a wide range in these time differences from a few seconds to several days, implying diverse browsing and decision-making styles. Further investigation could involve incorporating time-based features such as time since

last interaction, average time between events, or identifying bursty versus steady engagement patterns. These temporal signals may be key to understanding user intent and improving the accuracy of purchase classification.

These patterns confirm the feasibility of designing time-aware features such as

- Time between view, add_to_cart, and transaction
- Duration of engagement within a session
- Number and type of actions within certain time windows

Overall, the dataset is rich with behavioural signals and temporal structure, providing a solid foundation for addressing our research questions on predicting purchase behaviour.

ALGORITHM

Option 1: Gradient Boosted Decision Trees (GBDT) - from the textbook

Justification:
Gradient Boosted Decision Trees are a suitable candidate for this project because they can capture complex, non-linear patterns, which is ideal for our imbalanced dataset, as typical user behaviour leading to a purchase is rarely linear. For example, a user might view an item five times before making a purchase or buy it instantly. Additionally, tree-based models split data hierarchically and naturally account for interaction effects. GBDT is adept at handling heterogeneous feature types, processing both numeric and categorical features directly. Given the large-scale dataset of 2756101 records, GBDT is efficient without requiring extensive preprocessing. It supports missing values and does not necessitate normalization or scaling, as trees focus on relative feature values rather than absolute magnitudes. For feature importance and interpretability, GBDT offers tools like SHAP, which help visualize the influence of each behaviour on purchase predictions.

Option 2: Logistic Regression

Justification:
Logistic Regression is another good candidate for this project because it is an interpretable baseline for predicting the probability of a user making a purchase, which is easy to understand and communicate. The coefficients can effectively show how much each feature contributes to the model and how it increases or decreases the purchase likelihood. Besides, logistic regression models are fast to train, making them computationally efficient, even with large datasets using engineered features.