Predicting Purchase Intent of Customers for Targeted Marketing Using Clickstream Data

Submitted by:

Luz Joseph

Riken Kiri

Dan Nguyen

Ruhee Shrestha

Rama Shrotri

Instructor:

ZQ Cheng

Course:

MIS 587: Business Applications in Machine Learning
Date: May 6, 2025

Business Problem and Project Objectives

This project focuses on analyzing customer journey sequences from e-commerce clickstream data of customers in Indonesia to improve targeted advertising and enhance user experience. The goal is to determine whether a browsing session leads to a purchase, framing the problem as a binary classification task.

Research shows that personalized advertising based on clickstream behavior significantly increases conversion rates and revenue for online platforms. By examining temporal, sequential, and static features of user interactions, this analysis helps uncover behavioral patterns associated with purchases. For example, longer sessions that pass through deeper funnel stages such as ITEM_DETAIL → ADD_TO_CART are often indicative of stronger buying intent, whereas short, shallow sessions may suggest disinterest or poor site usability.

To evaluate which feature types best predict conversion, we compared models trained on static, temporal, and sequential features individually, and then all combined. Our objectives include:

- **Identify which clickstream features** (temporal, static, or sequential) best predict purchase conversions.
- **Evaluate the impact of combining** these feature types on model performance using AutoML (DataRobot).
- **Generate actionable insights** into user behavior patterns (e.g., session length, funnel depth) that correlate with conversion or drop-off.

By aligning these predictive insights with business goals, we can support personalized marketing strategies. For instance, high-probability leads can be targeted with tailored product recommendations or special promotions like free shipping. Similarly, warm leads or cart abandoners can be engaged through automated reminder emails and retargeted ads. These tactics help reduce churn and boost return on marketing investment.

Moreover, the implementation of a real-time predictive pipeline that integrates with CRM platforms allows for immediate activation of insights. This pipeline supports dynamic content delivery, UX optimization, and decision-making rooted in behavioral data. Ultimately, our project aims to bridge the gap between user interaction data and measurable business outcomes, enabling digital platforms to personalize user experiences, minimize friction, and increase sales through data-informed decisions.

Data Decisions

Our target variable is "converted", which we created when the session resulted in a successful purchase, derived from the transaction table where payment_status == 'Success'.

Feature Engineering Techniques

The first step we took for the clickstream, transaction and products data was to merge it and create new metrics for feature engineering to perform predictive modeling for finding purchase intent or user conversion in Data Robot/AutoML. Hence, our decision to have our target variable be "converted", which we created when the session resulted in a successful purchase, derived from the transaction table where payment_status == 'Success'.

To do this we transformed our raw dataset into a structure ready for predictive modeling in DataRobot. The final dataset combines temporal, statistical, and sequential features to enhance model performance on predicting user conversion. These features are as follows,

1. Temporal Features:

For temporal based feature modeling, we wanted to see the impact of user session-based data on the user conversion or purchase intent. Each action is encoded with a set of features that describe user interaction, such as event type (view, cart, etc.), product details before user makes by finding the total number of events, session duration, average session time to provide insights on overall engagement.

Features Created:

- **session_duration**: Total time (in seconds) between the first and last event in a session.
- **session_hour**: Hour of day when the first event in the session occurred (e.g., morning/evening browsing behavior).
- **event_per_sec**: Event frequency within a session, computed as num_events / session_duration.
- **events_per_session**: Normalized event intensity using num_events / (session_duration + 1).
- **avg_session_duration**: Average session time per user, derived from historical session data.
- **days_until_shipment**: Number of days between the session and the scheduled shipment date.
- **session_duration_bucket**: Binned version of session_duration:
  - <1 min, 1-3 min, 3-10 min, 10-30 min, 30+ min

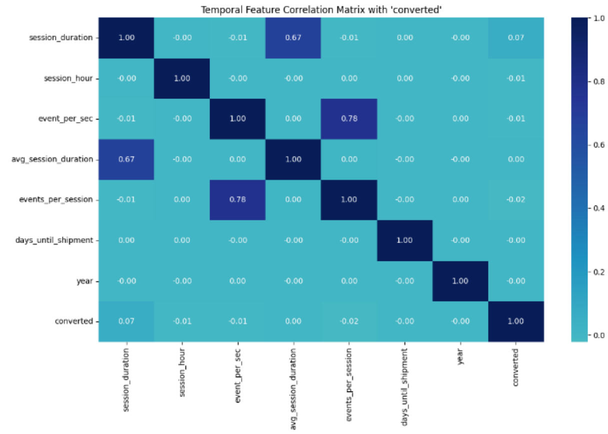


Figure 1: Correlation Matrix of Temporal Features

2. Static Features:

Static features are derived from stable, non-time-dependent attributes that describe either the user, the session, or the product in a summarized or categorical manner. They capture foundational context that can help characterize the overall profile of a user or interaction, such as customer demographic information.

Features Created:

- **promo_amount**: Promotional discount offered in the session.
- **total_amount_bucket**: Categorization of total transaction value:
    - Low, Medium, High, Very High, Premium
- **gender_encoded**: Gender transformed into binary format: F = 0, M = 1.
- **home_country, season, year**: Demographic and seasonal metadata.

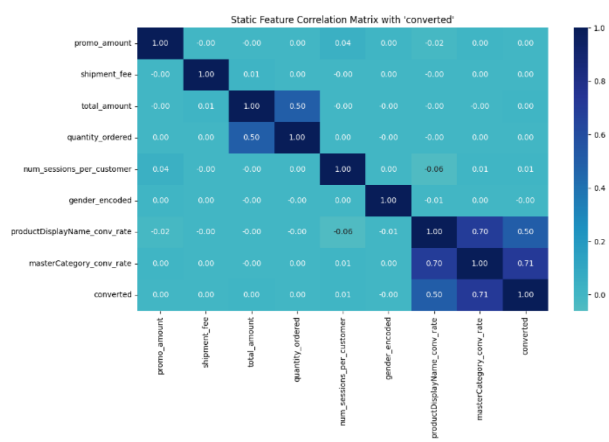


Figure 2: Correlation Matrix of Static Features

Conversion rates by category and product are strong predictors, as they carry historical success patterns. Features like promo_amount, shipment_fee, and total_amount have negligible direct correlation — these might influence behavior indirectly or need interaction terms, but we will still be using these variables for the model.

3. Sequential Features

Since the heart of clickstream data is, following a user's journey, we created a new feature 'funnel sequence', based on the event type journey in a session per customer, where we create a string capturing the order of event types for sequential modeling. We believed this feature would be useful for determining user drop off, and for applying deep learning which we hope to look at in future modeling.

Features Created:

- **funnel_sequence**: A string representing the ordered list of events (e.g., HOMEPAGE → SEARCH → ITEM_DETAIL).
- **reached_cart**: Binary indicator if ADD_TO_CART appeared in the funnel sequence.
- **funnel_depth_score**: Encoded funnel progress using event hierarchy:
  - ADD_TO_CART = 4
  - ITEM_DETAIL = 3
  - SEARCH = 2
  - HOMEPAGE = 1
  - Otherwise = 0
- **Event Counts (per session)**:
  - count_click,count_homepage,count_add_to_cart,count_item_detail, count_search, count_booking, count_promo_page, count_add_promo

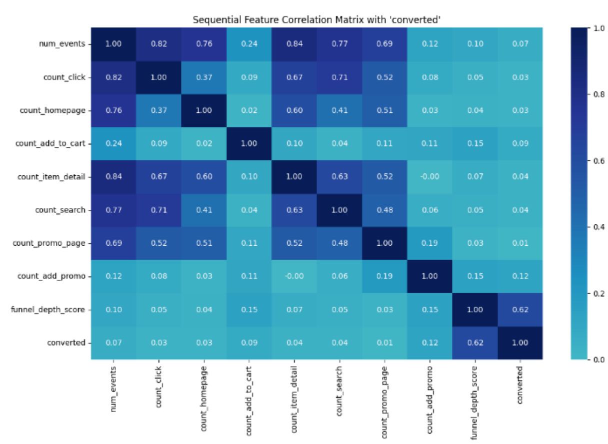


Figure 3: Correlation Matrix of Sequential Features

From the correlation matrix, funnel_depth_score is the most predictive — likely capturing the furthest user journey stage, making it an excellent behavioral summary. Similarly, the number of add promos leads to a purchase naturally.

Click counts and page views have weak correlation individually — suggesting that mere activity doesn't imply intent, but the nature of the action does (e.g., cart interaction vs homepage browsing). This suggests that depth over volume is more indicative of conversion.

	Median	Mean	Std Dev	Min	Max
session_duration	321978.0	950343.57	2314651.95	0.0	81620539.00
num_events	10.0	14.34	13.65	2.0	650.00
count_click	2.0	2.79	3.58	0.0	63.00
count_homepage	2.0	2.78	5.20	1.0	450.00
count_add_to_cart	1.0	2.16	2.20	0.0	52.00
count_item_detail	1.0	1.44	2.17	0.0	74.00
count_search	1.0	1.31	1.96	0.0	37.00
count_booking	1.0	0.95	0.21	0.0	1.00
count_promo_page	0.0	0.68	1.16	0.0	63.00
count_add_promo	0.0	0.36	0.48	0.0	1.00
session_hour	12.0	11.56	6.92	0.0	23.00
event_per_sec	0.0	0.00	0.24	0.0	152.76

Figure 4: Summary Statistics of Time-Based Aggregation of Sessions

The above summary statistics shows time-based aggregation per session per customer, where session duration is calculated as the time difference between the first and last event in a session.

The final dataset is a cleaned, merged, and enriched representation of user sessions, ready for supervised classification modeling in Data Robot. It combines time-aware behavioral metrics, sequential journey indicators, and all features were scaled, binned, or encoded appropriately for ingestion into AutoML/DataRobot .

Feature Selection

Feature selection was an integral step in optimizing the predictive power of the dataset while ensuring the integrity of the model training process.

Final Features to load in DataRobot:

Technique	Features / Strategy	Justification
Session Aggregation	session_duration, num_events, event_per_sec, events_per_session	Higher engagement often correlates with conversions
Funnel Progress Score	funnel_depth_score	Indicates user purchase intent based on behavior stage.
Temporal Features	session_hour, days_until_shipment	Reflect time-of-day behavior and urgency-driven purchases.
Conversion Rate Features	category_conv_rate, product_conv_rate	Useful generalization for sparse items; recomputed without label leakage.
Count-Based Behavior	count_add_to_cart, count_item_detail, count_search	Captures distinct types of user engagement.
Categorical Encoding	gender_encoded	Translates categorical values into numeric features suitable for ML models.

Experimentation in DataRobot

To evaluate the performance contribution of different feature types, we conducted an experiment by loading the full dataset into DataRobot and training separate models based on filtered subsets of features. These experiments allowed us to isolate and compare the predictive power of static, temporal, and sequential data representations.

- **Static-Only Model**: Used features such as gender_encoded, masterCategory_conv_rate, and productDisplayName_conv_rate. This model helped assess the baseline segmentation power of demographic and product-level data.
- **Temporal-Only Model**: Included session timing variables like session_duration, session_hour, event_per_sec, and days_until_shipment. These features captured contextual aspects such as user behavior during different times of day or urgency to convert.
- **Sequential-Only Model**: Focused on funnel_sequence, enabling DataRobot's internal deep learning models (e.g., GRUs or transformers) to interpret user navigation patterns.

Each model was evaluated for its precision, recall, and F1 score. The comparative results demonstrated that while temporal and static features offered complementary value, sequential features — especially event ordering captured in funnel_sequence — consistently delivered the highest performance. These findings validated our research-driven hypothesis and guided the decision to use a **sequential-only deep learning model** for final deployment.

Outcomes of Feature Type Modeling on Model performance

Adding feature-type-specific modeling significantly enhanced our analysis. It allowed us to isolate the individual and combined the impact of temporal, static, and sequential information on conversion prediction. This separation not only supported better model transparency but also helped confirm that no single type of feature dominates on its own. Instead, the interplay between behavioral depth (sequential), contextual timing (temporal), and categorical history (static) delivers the most robust and generalizable model. This layered insight is critical for both production deployment and ongoing feature strategy decisions. the performance contribution of different feature types, we conducted an experiment by loading the full dataset into DataRobot and training separate models based on filtered subsets of features. These experiments allowed us to isolate and compare the predictive power of static, temporal, and sequential data representations.

Model Selection Process:

Models Compared:

- **Light Gradient Boosting on ElasticNet Predictions (Best)**
- eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)
- RandomForest Classifier (Gini)
- Keras Slim Residual Neural Network Classifier using Training Schedule (1 Layer: 64 Units)

We compared multiple machine learning models using a consistent training dataset and evaluation framework, assessing their performance based on key classification metrics such as AUC, LogLoss, and F1-score. Each model was trained on the same feature set, with hyperparameters tuned via cross-validation or AutoML-guided optimization. We also evaluated training speed, model complexity, and interpretability to identify the most effective model for predicting purchase intent

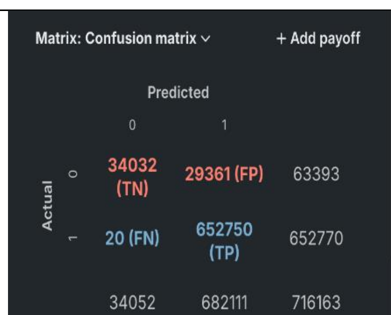
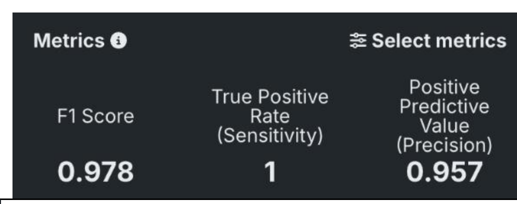
Model Performance Outcomes

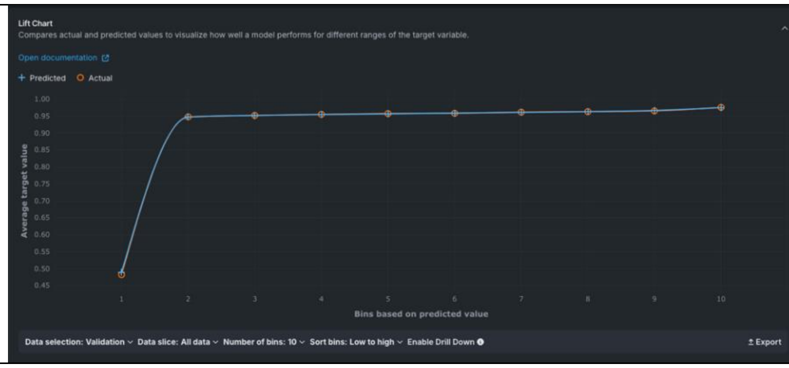
Each model was evaluated using LogLoss and AUC metrics across validation, cross-validation, and holdout sets:

- **Best Hybrid Model (All Features)**: Light Gradient Boosting on ElasticNet Predictions
  - **LogLoss (Holdout)**: **0.1655** (Lowest)
  - **AUC (Holdout)**: **0.8052** (Highest)
- **Best Static/Temporal Model**: eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)
  - **LogLoss (Holdout)**: 0.1675
  - **AUC (Holdout)**: 0.7841
- **Best Sequential Model**: Comparable results from: eXtreme Gradient Boosted Trees Classifier
  - **LogLoss (Holdout)**: ~0.2252
  - **AUC (Holdout)**: ~0.7312

These results confirm that combining all feature types in a hybrid model delivers the best predictive performance. However, the sequential-only model, despite slightly lower AUC, showed robust standalone performance and offers strong explainability for intent-focused use cases.

Model Quality Metrics on Hybrid Model

<table>
<tr><td>1. Confusion Matrix (CrossValidation) & (associated) Metrics</td><td>
</td><td>
</td></tr>
<tr><td>2. ROC Curve (CrossValidation)</td><td>
</td><td></td></tr>
</table>

<table>
<tr>
<td>3. Lift Chart</td>
<td>
</td>
</tr>
</table>

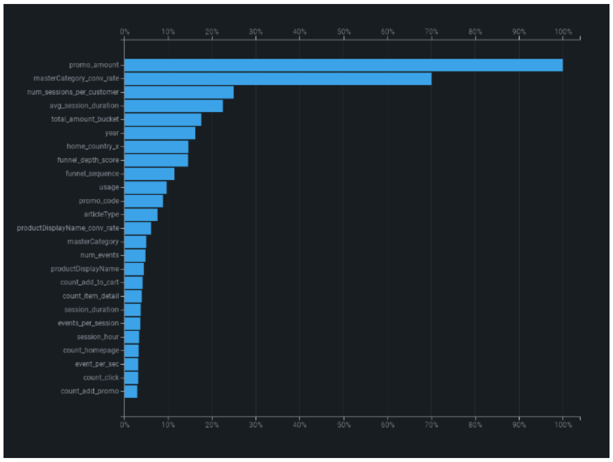
Based on the F1 score of 0.978, the model demonstrates strong overall performance with a good balance between precision and recall. The true positive rate is 1, which indicates that the model has high sensitivity for positive class and successfully identified all of the true positive cases. From the confusion matrix, there is less than 1% rate of the model predicting false positive and false negative. Additionally, because of the difference between the instances of negative and positive classes, there may be a class imbalance in the model.

For the ROC curves, the curve reaches the top-left corner, which indicates a high true positive rate and low false positive rate, adding up with the confusion matrix. On the other hand, the prediction distribution reveals a clear separation between classes due to the high probability of the model assigning to the positive and negative classes. Moreover, at the selected threshold of 0.6852, the model achieves minimal false positive rates. The threshold balances of precision and recall demonstrate a favor correct identification towards positives.

Finally, the lift chart confirms strong model calibration, as evidenced by the minimal gap between predicted and actual values across all bins. The steep initial rise in the chart indicates that high-confidence predictions are highly accurate. The plateau that follows suggests consistent performance across top probability ranges, further reinforcing the model's reliability.

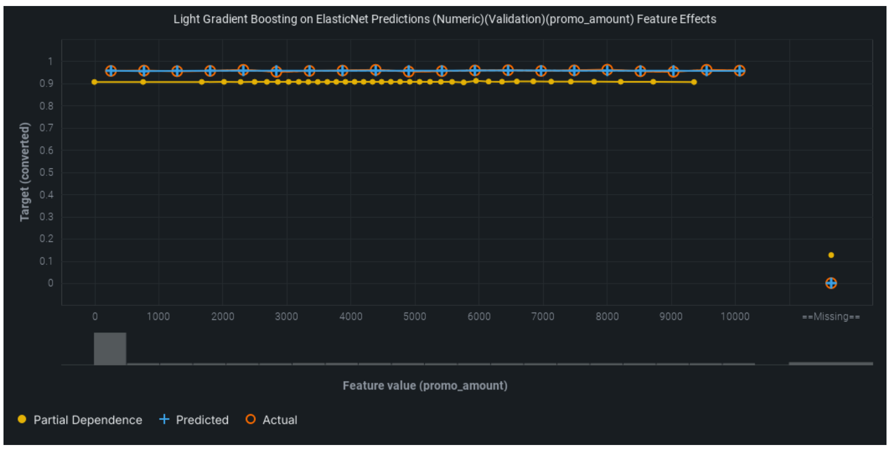
Most Predictive Features for Model Building

1. Hybrid Model Features (Static/Temporal/Sequential Features)
    a. **Feature Impact**



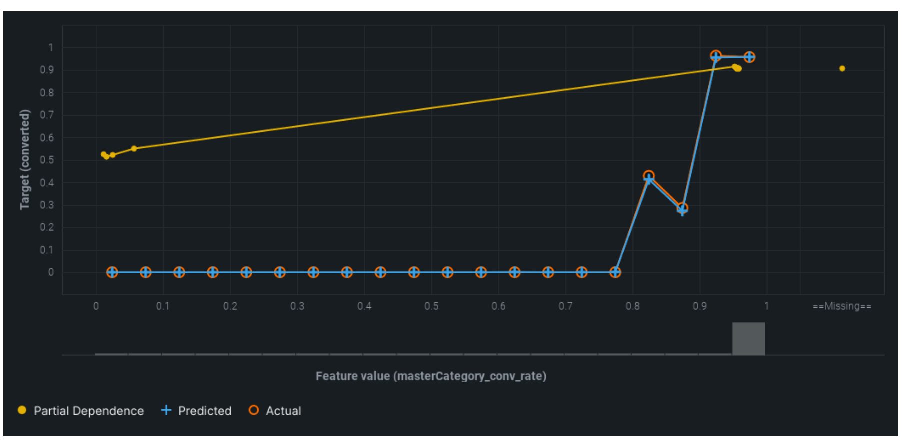
b. Feature Effects

i. Promo Amount



In the feature impact diagram above, the partial dependence (yellow line) and predictions (blue crosses) show very high and consistent conversion probabilities (~0.9) across most promo amounts. There is little variance in predicted impact, so the model sees nearly all non-missing promo amounts as highly indicative of purchase, suggesting that any promotion, regardless of size, significantly increases conversion likelihood.

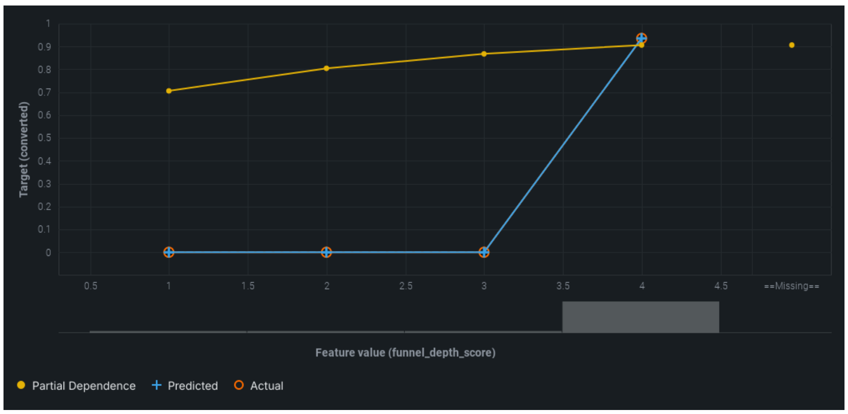
ii. Master Category Conversion Rate



In the above feature impact diagram of master category conversion rate, the partial dependence line (yellow) shows a steady increase in conversion likelihood as the category-level conversion rate increases. Predictions remain low and flat until the category conversion rate exceeds ~0.8, after which both predicted and actual conversions sharply increase.

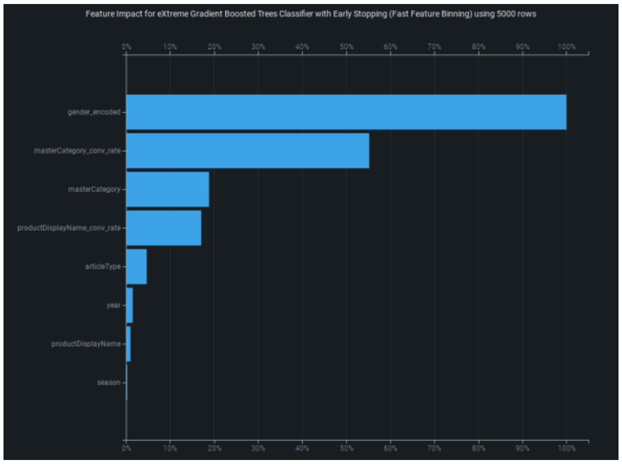
The model interprets categories with historically high conversion rates (e.g., >0.8) as strong signals of purchase intent. This reinforces the value of historical conversion patterns as a proxy for product appeal, but also highlights a potential sparsity issue, many categories may not reach the 0.8 mark, making this a strong but limited-use feature.

iii. Funnel Depth Score



For the funnel depth score, the partial dependence line shows a positive correlation between funnel depth and conversion probability. The model predicts zero conversion for scores ≤3, but sharp increase at depth 4, aligning with actual outcomes. Users reaching deeper funnel stages (score 4) are significantly more likely to convert. The flat prediction line for lower funnel scores suggests that the model treats shallow sessions as low-intent with high certainty. This behavior is expected and supports the idea that purchase intent is highly dependent on how far a user progresses through the shopping funnel, from browsing to checkout initiation.

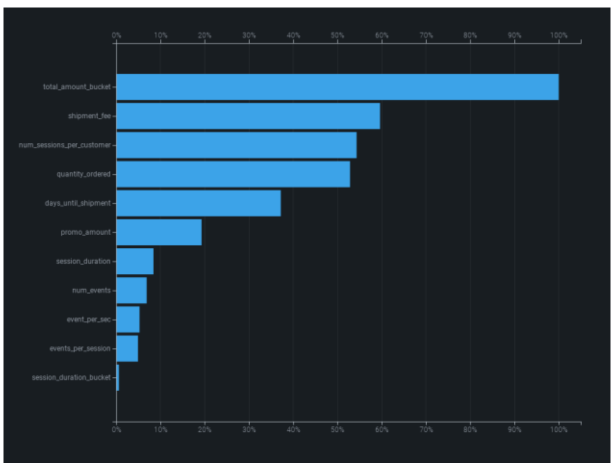
2. Static Features
    a. Feature Impact



Feature impact of Extreme Gradient Boost Tree Classifier Model – Static Features

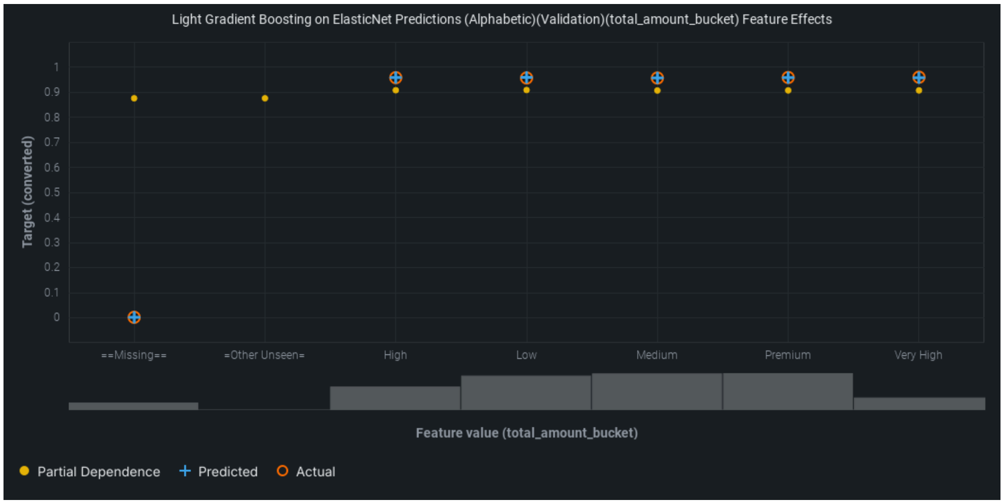
Gender and **historical conversion rates by category** are major drivers. Specific product types and seasonal factors are much weaker predictors in this dataset.

3. Temporal Features
    a. Feature Impact



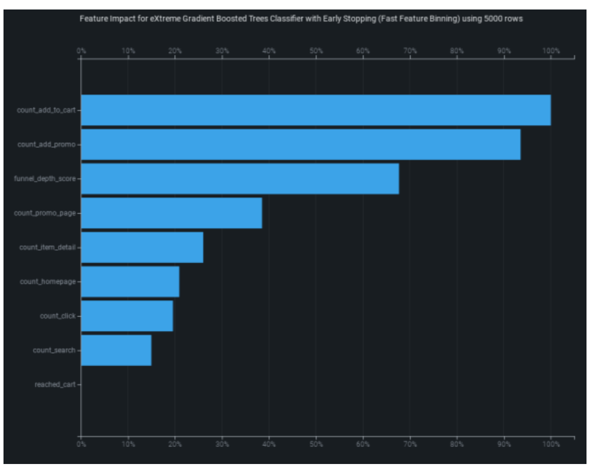
Feature impact of Extreme Gradient Boost Tree Classifier Model – Temporal Features

Purchase value (total_amount_bucket) dominates. **Loyalty (num_sessions)** and **order urgency (days_until_shipment)** help, but simple click counts and session time are **weak indicators** on their own.



For temporal features, total_amount_bucket, shipment_fee, and num_sessions_per_customer were highly impactful, suggesting that higher transaction values, loyalty, and shipping-related urgency are important contextual drivers. In contrast, features such as total event counts, session durations, and broad product or seasonal information showed comparatively weaker predictive power.

4. **Sequential Features**
   a. **Feature Impact**



Feature impact of Extreme Gradient Boost Tree Classifier Model – Sequential Features

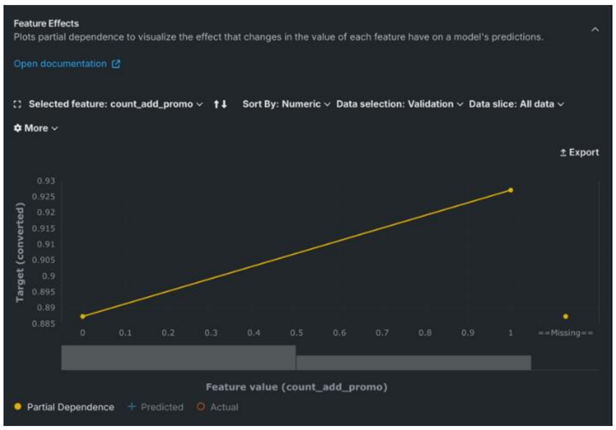
The sequential feature impact analysis reveals that actions directly tied to purchase intent, such as count_add_to_cart and count_add_promo, are the strongest predictors of user conversion. Users who frequently add items or promotions to their cart show the highest likelihood of completing a purchase.

The funnel_depth_score, representing how far users progress through the browsing journey, also plays a major role, reinforcing the idea that deeper engagement correlates with stronger buying intent.

Other features like count_promo_page and count_item_detail moderately contribute to predicting purchases, as they indicate a user's interest in specific items or discounts, though not as strongly as cart-related actions.

In contrast, features like count_homepage, count_click, and count_search show weaker impacts, suggesting that general browsing and searching activities alone are not sufficient indicators of purchasing behavior.

b. Feature Effects



Overall, the analysis emphasizes that behavioral depth, purchase value, and historical success signals are far more critical than general browsing activity or timing alone in predicting user purchase intent.

Areas where the Model Struggled

To better understand which types of user behavior were most predictive of purchase intent, we deliberately separated the dataset into three distinct modeling groups based on **feature type**: static, temporal, and sequential. The goal was to evaluate whether a particular feature class—such as time-based engagement patterns or sequential click behavior—would lead to a significantly stronger predictive signal. By training individual models on each feature set, we aimed to isolate the impact of that feature category and potentially identify one that would dominate in performance, offering more focused insights and optimization strategies.

However, despite this structured separation, we observed that **ensemble tree-based models such as XGBoost and LightGBM consistently outperformed others** across all feature groups. Regardless of whether the input features were temporal indicators (like session hour), user profile attributes (like gender or prior purchase history), or behavioral sequences (like funnel depth and click counts), these models demonstrated robust adaptability and strong generalization. Their ability to capture non-linear interactions and handle sparse or noisy tabular data made them more effective than deep learning models like Keras, which may require richer and more sequentially structured inputs to outperform.

One challenge with using features like session_duration, funnel_depth_score, and events_per_session is that they emphasize quantity over quality—long sessions or high click counts may reflect confusion or poor UX rather than true intent. Moreover, LightGBM model lacks native support for sequence modeling, so it overlooks the order and timing of user actions, which can be critical. For example, adding an item to the cart before viewing its details may indicate very different intent than following the usual flow. Aggregated metrics also blur behavioral differences, making it hard to distinguish quick, high-intent users from casual browsers. These limitations suggest potential gains from using sequence-aware models like RNNs or Transformers that can capture temporal dynamics and user journey patterns more effectively.

Dealing with Data Leakage

During the selection process, we identified two key sources of potential data leakage: productDisplayName_conv_rate and masterCategory_conv_rate. Initially, these conversion rate features were computed across the entire dataset, including future data not available at the time of each session. To mitigate this, we recalculated them in a time-aware manner, restricting the computation to only past data up to the point of each record.

We also flagged count_booking as a watchlist feature due to its potential to leak post-conversion behavior. A thorough audit confirmed that it was safely derived from pre-conversion interactions within each session.

Business Recommendations

To predict user purchase intent effectively, we built a model using clickstream session data, user demographics, and purchase history. Key features such as product preferences, interaction depth, and promo code availability emerged as strong predictors of conversion. With these predictive insights, marketing teams can segment customers based on their likelihood to convert and deliver personalized campaigns, including reminders to complete purchases, follow-up nudges, or dynamic promotional offers. This enables highly targeted interventions to improve conversion rates and customer engagement.

Our best-performing model, a Light Gradient Boosting Machine (LightGBM) trained on Elastic Net predictions, outperformed other contenders such as XGBoost, Keras deep learning models, and baseline classifiers within DataRobot. Several factors likely contributed to this model’s superior performance, particularly in the context of our clickstream dataset, which included a mix of static, temporal, and sequential features.

To support real-time engagement, we recommend implementing a **CRM framework** to centralize behavioral data, visualize funnel progression, and enable dynamic adjustments in marketing intensity based on conversion probability. Operational dashboards and predictive triggers can guide retargeting strategies, abandoned cart recovery, and promotional campaigns.

In addition to predictive modeling, insights on **product and category popularity** should be leveraged to optimize promotions and content. Popular items and conversion-driving product attributes (e.g., discounts, ratings, or availability of promo codes) can be highlighted to improve click-through and purchase rates. The analysis also supports broader marketing experimentation through A/B testing frameworks. By monitoring metrics like session length, bounce rates, and click-through rates, product and marketing teams can continuously refine messaging, layout, and flow.

Execution at Probability Thresholds

As detailed earlier of our distinctly defined feature-type grouping created, through static, temporal, and sequential; sequential- only based models gave greater insight into interpreting user navigation and user intent patterns. Thus, from the highest performing model on this feature type, we'd be able to predict the technical characteristics of customers lead to their conversion. Most specifically, pertaining to the actions to take in answer to the probability thresholds of these customer clusters. Concerning or hypothesis, we had three groups and their correlated probability threshold:

- **High Threshold (≥ 0.80) – "Hot Leads" -** Members captured through this threshold are those who are most likely to convert. With this reality, we would trigger a series of immediate personalized promotions and discounts, in the form of email/SMS offers, fast-lane checkout prompts, or loyalty incentives, to incentivized booking. Through these initiatives maximizing conversion of high-intent users, increase ROI on marketing spend.
- **Medium Threshold (0.50–0.79) – "Warm Leads" -** Members captured as 'Warm Leads," however, are those on the fence of conversions, likely needing further attention through campaigns or product recommendations. For this group we can have a continuous retargeting plan through with personalized ads, push notifications, or limited time offers.
- **Low Threshold (< 0.50) – "Cold Leads" - "**Cold Leads," however, are those within our customer segment whom we don't know much about. Members captured within this threshold can only be beneficially interacted with through improved user experience, strategy testing, collecting more data, and the implementation of surveys. All data collecting efforts should be implemented in ways without wasteful ad spending.

Implementation of our ML Model

Implementing a machine learning model within marketing operations begins with seamless integration into CRM platforms such as Salesforce or HubSpot. Our model predictions of conversion probabilities can be pushed into the CRM using APIs or ETL pipelines, where they populate custom fields associated with each user profile. This allows marketing and sales teams to access predictive insights in real time, directly within the tools they already use. These integrations form the backbone of real-time personalization and targeted outreach.

Once predictions are available in the CRM, teams can enable automated lead segmentation and prioritization. Users are classified into segments such as hot, warm, or cold leads based on probability thresholds output by the model. Hot leads, for example, can be fast-tracked to sales or offered incentives, while warm leads are placed into nurturing sequences, and cold leads receive low-touch engagement. This segmentation supports smarter allocation of resources and ensures the highest-potential users receive immediate attention.

CRM systems can trigger email, or ad retargeting campaigns using tools like journey builders, ensuring that outreach is timely and relevant. Meanwhile, marketing teams can monitor lead flow and conversion rates across segments through CRM dashboards, identifying friction points and areas for optimization.

Baseline and Profit Matrices (Payoff)

With the assumption that profit values depend on average transaction size, cost of marketing action, and predicted accuracy and that cost per promotion, or retargeting campaigns is lower than the average revenue from a converted user, the estimated financially implications of applying the models:

Prediction	**Customer Count**	**Possible Conversion Action**	**Corporate Action Taken**
High (≥ 0.80)	179041	Yes	Offer Discount
Medium (0.50 - 0.79)	268561	Maybe	Retarget
Low (< 0.50)	447601	No	Ignore

Prediction	**True Positive Rate (Assumed)**	**TP Profit**	**FP Cost**	**Payoff Estimate**
High (≥ 0.80)	90%	$80.5M	-$895k	High Profit (max ROI) - $79.6M
Medium (0.50 - 0.79)	60%	$80.5M	-$5.3M	Moderate Gain - $75.1M
Low (< 0.50)	20%	(opportunity cost)	0	Opportunity Cost - $44.7M

With the assumption that profit of a successful conversion is +$500, the cost of contacting the wrong person -$50, and no cost for not interacting with non-leads, we calculated the associated cost and profit of each probability threshold. Additionally, continuing with the assumption that of the "Hot Leads" group, 90% are actual leads; 60% of "Warm Leads" became actual customers; lastly, 20% of "Cold Leads" convert to real leads. From these assumptions we would see a payoff estimate of $79.6M, $75.1M and $44.7M opportunity cost, respectively, for our 895,203 customer points from our final dataset. With monitoring and maintenance, the organization would be able to recover the same respective costs from the false positives. These false positives being viewed as cost for the fact that focus on customers who aren't likely to convert leads to a waste of ad spending.

Final Recommendations for Implementation

Predicting purchase intent using machine learning (ML) models with real-time clickstream data provides businesses with a powerful advantage in understanding customer behavior, reducing cart abandonment, and executing targeted marketing campaigns. By analyzing behavioral signals such as session duration, page view sequences, product interactions, and promotional exposure, ML models can identify high-converting users in real time, enabling timely, personalized interventions.

To address this opportunity, we developed and evaluated multiple models and recommend deploying a Light Gradient Boosting Machine (LightGBM) stacked on Elastic Net predictions. LightGBM's ability to model complex, non-linear relationships, while leveraging Elastic Net's regularization strengths, makes it ideal for real-world, noisy behavioral datasets such as clickstream logs. Furthermore, it effectively integrates sequential journey patterns, demographic attributes, and temporal context to improve prediction accuracy.

Research by Bucklin and Sismeiro (2003) supports the underlying behavioral assumptions in our modeling. Their studies found that repeat visitors tend to view fewer pages without reducing time spent per page, indicating user learning and efficiency gains. These insights help us interpret high-conversion patterns in returning users as a sign of behavioral familiarity and focused intent. Additionally, the concept of "site lock-in", where deeper engagement increases likelihood of continued interaction aligns with our observation that longer or more immersive sessions are often precursors to purchase events.

However, we recommend implementing this model with safeguards, including post-deployment A/B testing to measure real-world ROI and ensure ad spending is yielding tangible conversion improvements. Ultimately, this ML-driven approach empowers organizations to shift from reactive to proactive marketing, delivering relevant, data-informed experiences that enhance user satisfaction, streamline campaign execution, and drive sustainable revenue growth.

Citations:

- A. Aylin Tokuç and T. Dag, "Predicting User Purchases From Clickstream Data: A Comparative Analysis of Clickstream Data Representations and Machine Learning Models," in IEEE Access, vol. 13, pp. 43796-43817, 2025, doi: 10.1109/ACCESS.2025.3548267.
- El Naqa, Issam, et al. "Artificial intelligence: A panacea for precision medicine." *Expert Systems with Applications*, vol. 141, 2020, 112901. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S0957417420301676.
- Tomescu, Dragos. "Conversion rate optimization in e-commerce: using machine learning to identify website satisfaction in clickstream patterns." Data Science & Society, June (2020).
- Bucklin, R. E., & Sismeiro, C. (2009). Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing. Journal of Interactive Marketing, 23(1), 35-48. https://doi.org/10.1016/j.intmar.2008.10.004 (Original work published 2009)