Churn Prediction using Machine Learning

Note: All the analysis and results of this project have been obtained using Python and the Jupyter Notebook has been provided for reference.

Understanding the problem:

Customer churn, which refers to users disengaging from a platform, poses a significant challenge for e-commerce businesses since it affects both revenue and growth. This project aims to create a strong churn prediction model by examining user behaviour, engagement levels, and purchasing habits. By utilizing insights from the research paper, the project integrates behavioural metrics, time-based trends, and machine learning techniques to pinpoint users at risk of churning and suggest practical interventions.

Data Inspection and Preprocessing:

1. **Data Loading and Overview**:

- The dataset was loaded, inspected, and summarized to understand its structure and content.
- Missing values were identified in category_code, brand, and user_session.

2. **Data Cleaning**:

- Duplicates were removed (655 entries).
- Missing values in category_code and brand were imputed with "unknown" to retain user-level context.
- Rows with null user_session (0.02%) were safely dropped.

3. **Data Type Corrections**:

- event_time was converted to datetime format to enable temporal feature extraction.

Exploratory Data Analysis (EDA):

Key Findings:

1. **Event Type Trends:**
    - Most interactions were "view" events, with significantly fewer "cart" and "purchase" events.
    - Churn users showed lower engagement with "view" events compared to no-churn users.
2. **Category and Brand Analysis:**
    - Categories like stationery.cartridge exhibited higher churn rates, indicating potential dissatisfaction.
    - Brands with high churn rates (e.g., 45 out of 50 brands having a churn rate of 1.0) suggest the need for targeted retention strategies.
3. **Temporal Patterns:**
    - Activity was highest during weekdays, with Monday being the peak day.
    - Event counts dropped on weekends, with Saturday having the lowest engagement.
4. **Price Analysis:**
    - Prices were skewed; Box-Cox transformation and winsorization improved distribution.
    - Median prices were slightly lower for churn users, indicating possible price sensitivity.
5. **Time of Day:**
    - Engagement peaked between morning and midday, declining steadily after 3 PM. Churn users showed lower activity during these peak times.

Defining Churn:

Definition: A user is considered churned if they meet one or both of the following:

1. **No Purchase Activity**: No purchase in the last 30 days.
2. **Overall Inactivity**: No activity (view, cart, or purchase) in the last 90 days.

Rationale:

- Aligns with typical e-commerce purchasing cycles.
- Combines purchase and engagement thresholds for comprehensive identification of at-risk users.

```
Churn Distribution:
churn
0    216726
1    190511
```

0 indicates ‘no churn’ and 1 indicates ‘churn’.

Feature Engineering:

Before feature engineering, data pre processing was carried out in which the categorical and temporal features were encoded using suitable methods.

Features Created:

1. **User-Level Metrics**:
    - total_events, total_views, total_purchases, avg_session_duration, last_event_time, recency_days, and purchase_to_view_ratio highlight user activity and intent.
    - Top viewed and purchased categories identify user preferences for personalized interventions.
2. **Session-Level Metrics**:
    - session_length, session_events, total_sessions, avg_session_length and bounce_rate measure engagement depth and interaction quality.
3. **Temporal Features**:
    - event_day, event_hour, and event_month capture seasonal and hourly trends.

```
0   event_time              884312 non-null  datetime64[ns, UTC]
1   product_id              884312 non-null  int64
2   category_id             884312 non-null  int64
3   category_code           884312 non-null  object
4   brand                   884312 non-null  object
5   price                   884312 non-null  float64
6   user_id                 884312 non-null  int64
7   user_session            884312 non-null  object
8   churn                   884312 non-null  int64
9   hour                    884312 non-null  int32
10  day_of_week             884312 non-null  object
11  brand_encoded           884312 non-null  int64
12  event_type_purchase     884312 non-null  bool
13  event_type_view         884312 non-null  bool
14  event_day               884312 non-null  int32
15  event_hour              884312 non-null  int32
16  event_month             884312 non-null  int32
17  category_encoded        884312 non-null  int64
18  session_length          884312 non-null  float64
19  session_events          884312 non-null  int64
20  total_events            884312 non-null  int64
21  total_views             884312 non-null  int64
22  total_purchases         884312 non-null  int64
23  avg_session_duration    884312 non-null  float64
24  last_event_time         884312 non-null  datetime64[ns, UTC]
25  purchase_to_view_ratio  884312 non-null  float64
26  recency_days            884312 non-null  int64
27  top_viewed_category     884312 non-null  object
28  top_purchased_category  186897 non-null  object
29  total_sessions          884312 non-null  int64
30  avg_session_length      884312 non-null  float64
31  bounce_rate             884312 non-null  float64
```

The imbalance of classes was handled using SMOTE. The feature importance of various features was determined using Random Forest.

Model Building and Evaluation:

Models Used:

1. **XGBoost**:

    - Achieved higher recall (99.18%) and F1 score (99.45%), excelling in identifying at-risk users.
    - Fewer misclassifications (1059 false negatives), making it effective in minimizing missed churn cases.

XGBoost Model Performance Metrics:

	Accuracy	Precision	Recall	F1 Score	Confusion Matrix	AUROC Score
XGBoost	0.994636	0.997168	0.991806	0.994480	[[135685, 364], [1059, 128186]]	0.999927

2. **LightGBM**:

    - Delivered slightly better precision (99.76%), reducing false alarms.
    - Performed consistently across folds but with slightly higher misclassifications (1306 false negatives).

	Accuracy	Precision	Recall	F1 Score	Confusion Matrix	ROC AUC
LightGBM	0.993920	0.997606	0.989895	0.993736	[[135742, 307], [1306, 127939]]	0.999920

Cross-Validation:

- XGBoost had higher mean accuracy (99.48%) with lower variability, indicating consistent performance. Hence, we select XGBoost as our final model.

Interpretation & Recommendations:

Key Insights:

1. **Critical Predictors of Churn**:

    - **recency_days**: This feature indicates how recently a user interacted with the platform. Users with high recency values (longer gaps since their last interaction) are more likely to churn. This aligns with the principle that inactivity is a strong indicator of disengagement.

- **purchase_to_view_ratio**: This feature reflects the conversion efficiency, i.e., how effectively users transition from browsing to purchasing. A low ratio signals users who engage in exploratory behavior without making purchases, suggesting potential dissatisfaction or lack of intent to buy.
- **total_purchases**: Historical spending patterns are indicative of user loyalty. Users with low purchase counts may not be deriving enough value from the platform, making them more prone to churn.

2. **Category and Brand-Level Churn**:

    - Certain categories, such as stationery.cartridge, have disproportionately high churn rates, suggesting dissatisfaction or a mismatch between user expectations and product offerings.
    - Brands with high churn rates (e.g., 45 of the top 50 brands) suggest potential gaps in product quality, pricing, or post-purchase experience.

Actionable Steps:

1. **Personalized Retention Campaigns**:

    - **Target Group**: Users with high recency_days and low purchase_to_view_ratio.
    - **Strategies**:
        - **Re-engagement Emails**: Send personalized emails highlighting products they've viewed but not purchased, along with limited-time discounts or recommendations based on their browsing history.
        - **Loyalty Rewards**: Offer loyalty points or discounts for the next purchase to encourage buying behavior.
        - **Feedback Mechanisms**: Prompt users to provide feedback on why they haven't completed purchases, giving insights into pain points (e.g., pricing, availability, usability).

2. **Category-Specific Retention Strategies**:
    - **High-Churn Categories**:
        - **Stationery.cartridge, electronics.telephone, computers.peripherals.printer:** Investigate why these categories has higher churn. Potential actions include price adjustments, bundling offers, or improving product descriptions to better set customer expectations.
    - **Stable Categories**:
        - Categories such as computers.components.videocards show a higher proportion of no-churn users. Promote these categories to at-risk users to boost engagement with products that have demonstrated retention potential.

3. **Peak Time Engagement**:
    - **Morning to Midday Focus**: Activity peaks during these hours, indicating when users are most engaged. Use this time window to:
        - Launch time-sensitive offers, such as flash sales or limited-time discounts.
        - Schedule push notifications or email campaigns to coincide with these hours for maximum visibility and conversion.
4. **Brand Retention Strategies**:
    - For brands with a churn rate of 1.0, conduct an in-depth analysis to identify reasons for dissatisfaction. Actions might include:
        - Collaborating with these brands to improve product quality or resolve common complaints.
        - Providing stronger return/refund policies or post-purchase support to enhance user trust.
5. **Dynamic Pricing Strategies**:
    - **Observation**: Churn users tend to make purchases at slightly lower median prices compared to no-churn users. Implement personalized pricing strategies, such as offering discounts or exclusive deals to users showing churn tendencies.
6. **Seasonal Campaigns**:
    - Utilize the event_month feature to identify seasonal trends and develop campaigns tailored to those months. For example:
        - Promote school supplies during back-to-school months.
        - Highlight gifts and festive items during the holiday season.
7. **Session-Based Interventions**:
    - **Bounce Rate**: Users with high bounce rates (single-event sessions) could be shown targeted exit-intent pop-ups offering discounts or relevant product recommendations.
    - **Session Duration**: Identify users with decreasing session lengths and send personalized suggestions to rekindle interest, such as curated product lists or tutorials.
8. **Enhancing Post-Purchase Experience**:
    - For churn-prone brands or categories, focus on improving post-purchase interactions:
        - Send follow-up emails for reviews or feedback.
        - Provide tutorials or setup guides for complex products (e.g., electronics).

Conclusion:

By utilizing insights from factors such as recency_days, purchase_to_view_ratio, and churn trends within specific categories and brands, these strategies can effectively minimize churn and improve customer retention. Merging targeted interventions, dynamic pricing, and increased engagement during peak periods creates a more tailored and efficient retention strategy, aligning with both business goals and customer satisfaction.

By:
Name: Tarun Kumar Behera
Email: tkbehera.work304@gmail.com