User Churn Prediction

REPORT

by

Sasi Venkat Gowd Dasari

1. Introduction:

This report outlines the approach taken to predict user churn in an e-commerce environment using a dataset of user events. The analysis includes data preprocessing, exploratory data analysis (EDA), model training, and actionable recommendations aimed at reducing churn.

2. Data Preprocessing :

- **Dataset Overview:-** The dataset consists of 885,129 entries and 9 columns:
  - **event_time:** Timestamp of the event.
  - **event_type:** Type of event (e.g., view, cart, purchase).
  - **product_id:** Unique identifier for the product.
  - **category_id:** Unique identifier for the product category.
  - **category_code:** Hierarchical category code.
  - **brand:** Brand of the product.
  - **price:** Price of the product.
  - **user_id:** Unique identifier for the user.
  - **user_session:** Identifier for the user session.
- **Handling Missing Values**:-The dataset contains missing values, particularly in:
  - **category_code:** 26.7% missing.
  - **brand:** 23.9% missing.
- **Duplicate Rows:-** A total of 655 duplicate rows were identified and removed, resulting in 884,474 entries.
- **Null Value Removal:-** Rows with null values were dropped, leading to a total null value percentage of approximately 5.63%.
- **Data Type Conversion:-** The event_time column was converted to a datetime format to facilitate time-based analysis.

3. Exploratory Data Analysis (EDA):

EDA is a critical step in understanding the dataset and uncovering patterns that can inform model development. The following analyses were performed:

Event Type Distribution:

A time series plot was created to visualize the distribution of different event types (view, cart, purchase) over time. This helps in understanding user engagement trends and identifying peak activity periods.

Popularity Analysis:

Brand Popularity: A count plot was generated to identify the top 10 popular brands based on user interactions. This analysis highlights which brands are most frequently viewed or purchased, providing insights into user preferences.

Category Popularity: Similar to brand analysis, a count plot was created for product categories to identify the top 10 popular categories. This helps in understanding which product categories attract the most user interest.

User-Level Summaries:

A summary of user behavior was created by aggregating data based on user_id. Key metrics calculated include:

Total Spent: The total amount spent by each user.
Average Spent: The average spending per event for each user.
Event Count: The total number of events (views, carts, purchases) for each user.
Active Days: The number of unique days a user has interacted with the platform.
View-to-Purchase Ratio: This metric indicates how many times a user viewed products compared to how many purchases they made, providing insights into user engagement and conversion rates.

Churn Definition

Churn was defined as inactivity for 30+ days. This threshold is commonly used in e-commerce to identify users who may be at risk of leaving the platform.

4. Model Training:

Feature Scaling:

Key features, including total_spent and avg_spent, were standardized using a StandardScaler to ensure that all features contribute equally to the model training process.

Model Selection: Two machine learning models were trained:

Random Forest Classifier:

A robust ensemble method that can handle non-linear relationships and interactions between features.

Logistic Regression:

A simpler model that provides interpretable results and is effective for binary classification tasks.

Training and Testing:

The dataset was split into training and testing sets using an 80/20 split. This allows for model evaluation on unseen data.

5. Final Scores and Performance Metrics:

Random Forest Classifier:

Accuracy: 77.34%

Precision:

Churned Users: 0.38
Non-Churned Users: 0.80

Recall:

Churned Users: 0.13
Non-Churned Users: 0.94

F1-Score:

Churned Users: 0.20
Non-Churned Users: 0.87

Logistic Regression:

Accuracy: 79.21%

Precision:

Churned Users: 0.62
Non-Churned Users: 0.79

Recall:

Churned Users: 0.01
Non-Churned Users: 1.00

F1-Score:

Churned Users: 0.03
Non-Churned Users: 0.88

6. Interpretation & Explanation: Model Performance Insights.

Random Forest Classifier:

The model shows a good ability to identify non-churned users (high recall of 0.94),

indicating that it is effective at predicting users who are likely to remain active. However, the low precision for churned users (0.38) suggests that it misclassifies many non-churned users as churned.

The F1-score for churned users (0.20) indicates that the model struggles to accurately predict churned users, which is critical for retention strategies.

Logistic Regression:

This model achieved a higher overall accuracy (79.21%) compared to the Random Forest model. However, it also shows a low recall for churned users (0.01), indicating that it fails to identify most of the users who actually churned.

The precision for churned users (0.62) is better than that of the Random Forest model, suggesting that when it predicts a user will churn, it is more likely to be correct.

Key Takeaways:

Both models demonstrate strengths and weaknesses in predicting user churn. The Random Forest model is better at identifying non-churned users, while the Logistic Regression model has a higher overall accuracy but struggles with churn predictions.

The low recall for churned users across both models highlights the need for further refinement in the predictive capabilities, particularly in identifying users at risk of churning.

7. Recommendations:

Based on the analysis and model performance, the following recommendations are proposed to reduce user churn:

Target Inactive Users:

Implement automated reminders for users who have not interacted with the platform for 15+ days. Personalized messages can encourage users to return.

Focus on High Spenders at Risk:

Identify high-spending users who show declining activity. Offer exclusive discounts or personalized product recommendations to re-engage them.

Convert Browsers into Buyers:

Develop cart abandonment campaigns to follow up with users who leave items in their carts. Consider offering limited-time discounts to create urgency.

Re-Engage Long-Time Users:

Celebrate user milestones (e.g., anniversaries) with personalized offers to encourage continued engagement.

Monitor and Improve User Experience:

Collect feedback from churned or inactive users to understand pain points. Use this information to optimize the user journey and enhance satisfaction.

Proactive Communication:

Utilize chat support to assist users during their browsing experience. Notify users about relevant product restocks or price drops to keep them engaged.

8. Conclusion:

The analysis successfully identified key features influencing user churn and implemented machine learning models to predict churn with reasonable accuracy. The Logistic Regression model outperformed the Random Forest model in terms of overall accuracy, but both models exhibited challenges in accurately predicting churned users.

By leveraging insights from user behavior and implementing the recommended strategies, the e-commerce platform can enhance user retention, improve satisfaction, and ultimately boost revenue. The findings underscore the importance of continuous monitoring and adaptation of strategies to effectively address user churn in a dynamic market environment.