Predicting Customer Purchasing Behavior in E-commerce Using Machine Learning

Yifan (Crystal) Cai

Siwei (David) Ran

Sanjana Kallol

2023/03/20

1. Executive Summary:

In this project, we aimed to develop predictive models that accurately determine whether a customer is likely to complete a purchase after adding a product to their cart in an e-commerce setting. To achieve this, we employed three classification models, including logistical regression, random forest, and neural network, and evaluated their performance to gain valuable insights into customer purchasing behavior. Our analysis was based on e-commerce customer behavior data from a large multi-category online store that was collected by the Open CDP project [1].

As the original data spanned seven months (from October 2019 to April 2020) and was too extensive for analysis, we randomly sampled one million observations from each month to create a new dataset. We then processed the features in the dataset by removing duplicate entries, irrelevant columns, and missing data, ensuring that the data was clean and suitable for analysis. Next, we conducted exploratory analysis of the features of the data based on dimensions of customer, products, and events. In total, we engineered eight variables as inputs for our three models.

After evaluating the performance of the models and making comparisons on the validation set, we found that logistical regression worked best in predicting customer purchase behavior compared with the other two models. Moreover, we identified that the activity_count variable was the most critical factor in predicting customer purchases. This implies that customers who perform more activities during a session have a higher likelihood of completing a purchase.

Our findings provide valuable insights into customer purchasing behavior in the e-commerce industry, allowing businesses to tailor their sales strategies and offer personalized incentives to encourage customers to complete their purchases. By leveraging these insights, businesses can make data-driven decisions that increase sales and revenue while also enhancing the overall shopping experience for customers.

2. Background, Context, and Domain Knowledge:

In recent years, the e-commerce industry has undergone substantial growth, which has been further accelerated by the COVID-19 pandemic, making online shopping an essential part of our daily lives. As a result, the global e-commerce sales are projected to reach over $6.3 trillion in 2023 and are expected to continue to grow, with sales forecasted to reach $8 trillion by 2026[2]. Despite this growth, businesses still face various challenges, with one of the most critical being the ability to predict customer behavior accurately.

When a customer adds a product to their cart, it is not a guarantee that they will complete the purchase, resulting in lost sales and revenue for businesses. Therefore, it is crucial to understand why customers abandon their carts and how to incentivize them to complete their purchases. By accurately predicting customer behavior, businesses can tailor their sales strategies and offer personalized incentives to encourage customers to complete their purchases. For instance, if a particular customer is predicted to be unlikely to complete their purchase, the business can offer a targeted discount or promotion to incentivize them to do so.

Machine learning models offer a promising solution to analyze vast amounts of customer behavior data and identify patterns that can help businesses understand what drives customers to complete their purchases. This includes factors such as the customer's browsing history, time spent on the website, type of product added to the cart, and demographics. Ultimately, accurate prediction of customer behavior is critical for businesses in the e-commerce industry as it can help them optimize their sales strategies, increase revenue, and enhance the overall shopping experience for their customers.

3. Business Problem and Strategy

The world of e-commerce has long been striving to find ways to encourage customers to complete their purchases. Methods such as email marketing, retargeting ads, and offering discounts and promotions have been commonly used, but their effectiveness may be limited as they are based on assumptions about customer behavior. In light of this, we propose a new approach that leverages machine learning models to analyze customer behavior data and provide insights into what motivates customers to complete their purchases.

Our goal is to develop predictive models that can accurately determine whether a customer is likely to make a purchase after adding a product to their cart. This data-driven approach empowers businesses to make informed decisions about when and how to incentivize customers to complete their purchases, thus improving sales and revenue. Furthermore, this approach helps to enhance the overall shopping experience by providing businesses with insights that can be used to optimize their sales strategies and improve customer satisfaction.

Using machine learning to analyze customer behavior data is aligned with the modern business model of making data-driven decisions. By accurately predicting customer behavior, businesses can reduce the resources and time wasted on ineffective sales strategies, resulting in increased efficiency and cost savings. Ultimately, our approach seeks to revolutionize the e-commerce industry by offering a new, smarter way to encourage customers to complete their purchases.

4. Data Analysis

4.1 Data Sources and Dataset Description

The e-commerce customer dataset contains a wealth of information about user events, including product views, cart additions, and purchases. In its original form, the dataset comprised approximately 70 million observations for each month between October 2019 and April 2020, stored in a flat file format with each row representing a unique event. The columns of the dataset provided valuable insights into user behavior, including user_id, product_id, category_code_level1, category_code_level2, event_type, event_time, original_price, discounted_price, and brand.

To make our analysis more manageable, we randomly sampled one million observations from each month of the original dataset and combined them to create a new dataset for our analysis. The resulting dataset contained eight variables, with detailed descriptions of each variable provided in Table 01. This approach allowed us to effectively analyze the data and gain valuable insights into customer behavior.

Table 01 : Description of Variables in the Dataset

Property	Description
event_time	Time when event happened at (in UTC).
event_type	Only one kind of event: purchase.

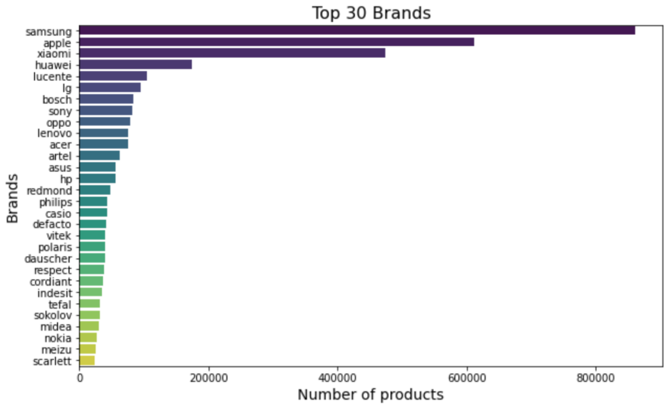
product_id	ID of a product
category_id	Product's category ID
category_code	Product's category taxonomy (code name) if it was possible to make it.
brand	Downcased string of brand name. Can be missed.
price	Float price of a product. Present.
user_id	Permanent user ID.
user_session	Temporary user's session ID.

4.2 Exploratory Data Analysis

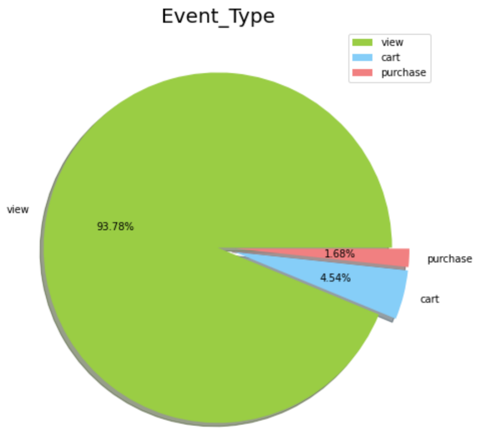
Exploratory data analysis (EDA) is a crucial step in gaining insights and understanding from a dataset without making any assumptions or predictions. In this case, the EDA aims to present an overview of the dataset, followed by a deeper look into products, users, and events.

After conducting exploratory data analysis (EDA) on the provided dataset, several interesting findings emerged. The dataset contains 7 million rows of data, with 3163042 unique visitors and 244624 unique products. The dataset spans over several months, with 1 million data samples from each month. When it comes to the top brands, the dataset includes several brands, but the top 30 brands hold the majority of the products. We can aslo tell that Samsung has an extremely high proportion of products compared with all the other brands. Other brands in the top 30 include a mix of well-known and lesser-known brands, with some specializing in certain categories of products (see plot 01).

Plot 01 : Top 30 brands



Plot 02: Top 30 brands



The three types of events present in the dataset are view, cart, and purchase. The majority of events are views, accounting for 93.3% of all events, followed by cart events (4.5%), and purchase events (1.7%) (see plot 02). This suggests that most users are browsing the website but not necessarily making purchases. This insight can be used to improve the website and increase conversions by identifying potential barriers to purchasing and optimizing the user experience.

A deeper look into the products displays the product ID, name, and price, which can help identify the most profitable products.

Next, identifying the top 10 products by earnings and the top 10 products by earnings per session helps us determine which products generate the most revenue for the company and are most efficient at converting traffic into purchases. By taking the conversion rate into account, the analysis can provide more accurate insights into which products are the most profitable (check code).

Lastly, the analysis identifies the top 10 products with the highest conversion rates. These products are most successful at turning website visitors into customers, and their success can guide business decisions on optimizing the website's design and functionality to improve the conversion rates of less successful products. Furthermore, analyzing user behavior can provide insight into the effectiveness of marketing campaigns, and businesses can use these insights to create better customer experiences (check codes).

4.3 Feature Engineering

The feature engineering process involves creating new variables and transforming existing ones to prepare a dataset used for modeling.

Firstly, the dataset is filtered to only include events with event types of "cart" and "purchase" and then dropping any duplicates. This is done to focus on the events that represent a user's intent to purchase a product. A new column "is_purchased" is then created to indicate whether a user eventually made a purchase. This column is set to 1 if the event type is "purchase", and 0 otherwise. Next, the "is_purchased" column is grouped by user_session and product_id and the maximum value is taken, indicating whether a user made a purchase for a particular product during the session.

The dataset is then further filtered to only include events with event type "cart" and dropping any duplicates for user_session, product_id, and is_purchased columns. This step ensures that each row in the dataset represents a unique session and product combination. The weekday of the event_time column is extracted using datetime and stored in a new column "event_weekday". This variable can be useful in identifying whether there are any specific weekdays that have higher purchase rates compared to others.

The category_code variable is split into two new variables, "category_code_level1" and "category_code_level2", which represent the first and second level of the category hierarchy respectively. This can help in identifying which categories of products are more popular among users.

The next step involves identifying the users who have engaged in at least one cart or purchase event and creating a new dataset "cart_purchase_users". This is done to focus on the users who have demonstrated an intent to purchase and have a higher likelihood of converting. The dataset is further filtered to remove any rows with missing values.

A new dataset "cart_purchase_users_all_activity" is then created to include all the events performed by the users in the "cart_purchase_users" dataset. This can provide additional context on the users' behavior and preferences.

Finally, the number of events performed in each user_session is calculated and stored in a new column "activity_count". This variable can be useful in identifying whether users who perform more activities during a session have a higher likelihood of making a purchase.

4.4 Model Building

We then developed a binary classification model using three different algorithms: Logistic Regression, Random Forest, and Neural Network. And we wanted to compare the performances of three models using metrics like accuracy, precision, recall, and F-beta scores.

Before building the model, we resampled data because purchased and non-purchased values were not evenly distributed, or extremely imbalanced. Down-sampling was used to balance the data, which involved randomly removing observations from the majority class to prevent the model from being biased towards the majority class.

The input features of the model were: brand, price, event_weekday, category_code_level1, category_code_level2, and activity_count. These variables were encoded using the LabelEncoder function from the sklearn library. Finally, the data was split into training and testing sets using the train_test_split function from the sklearn library.

4.5 Model Performances and Interpretations

The results of performance of three models can be found in Table 02. The Logistic Regression model has the highest accuracy score of 92.21%, followed closely by the Random Forest model with an accuracy score of 90.27%. The Neural Network model has the lowest accuracy score of 89.62%. This means that the Logistic Regression model performs better than the other models in terms of overall performance. When looking at the Precision score, the Logistic Regression model again outperforms the other models with a score of 88.44%. The Random Forest model has a precision score of 87.65%, and the Neural Network model has a precision score of 87.92%. In terms of Recall, the Logistic Regression model has the highest score of 97.66%, indicating that this model is better at identifying true positives. The Random Forest model has a recall score of 94.42%, and the Neural Network model has a recall score of 92.37%. The Logistic Regression model has the highest F-beta Score of 92.42%, indicating the best balance between precision and recall. The Random Forest model has an F-beta Score of 90.36%, and the Neural Network model has an F-beta Score of 89.78%.

Overall, based on the evaluation metrics used, the Logistic Regression model is the best-performing model in this analysis, with the highest accuracy, precision, recall, and F-beta Score.

Table 02: Performances of Models

Model	Accuracy	Precision	Recall	F-beta Score
Logistic Regression	92.21%	88.44%	97.66%	92.42%

Random Forest	90.27%	87.65%	94.42%	90.36%
Neural Network	89.62%	87.92%	92.37%	89.78%

4.6 Importances of Variables

The importance of variables can give insights into which features have the strongest influence on the target variable in a machine learning model. In this case, we have the importances of variables from both logistic regression and random forest models.

For the logistic regression model, the most important feature is "activity_count" with a score of 3.79080. This means that for every unit increase in "activity_count," the odds of a positive outcome for the target variable increase by a factor of e^3.79080, or about 44 times. The second most important feature is "category_code_level1" with a score of 0.02087, indicating that it has a much smaller effect on the target variable compared to "activity_count." The remaining features have scores that are close to 0, which means that they have very little influence on the target variable.

For the random forest model, the most important feature is "activity_count" with a score of 0.742102, followed by "price" with a score of 0.139279. This suggests that "activity_count" is a very strong predictor of the target variable in both models. The remaining features have scores that are much smaller, indicating that they have less influence on the target variable.

It is interesting to note that the rankings of the features are somewhat different between the two models. This is not surprising, as different models may have different ways of weighing the importance of features. However, it is reassuring that "activity_count" is consistently ranked as the most important feature in both models.

In conclusion, the results suggest that "activity_count" is the most important predictor of the target variable in both the logistic regression and random forest models. This variable could be a key factor to consider when making predictions or trying to understand the underlying factors that contribute to the target variable.

5. Recommendations and Business Value

Based on our analysis, we recommend that businesses in the e-commerce industry use logistic regression as the primary model for predicting customer behavior. Our analysis found that logistic regression outperformed both random forest and neural network models in predicting customer behavior. Additionally, we recommend that businesses focus on the activity count feature when predicting customer behavior. This feature has the highest importance in predicting whether a customer will make a purchase after adding a product to their cart. Therefore, businesses should consider offering targeted incentives to customers who engage in more activities during their sessions, such as offering a discount or promotion to customers who spend more time browsing or adding multiple products to their cart.

The business value of accurately predicting customer behavior is significant. By using predictive models, businesses can optimize their sales strategies and offer personalized incentives to encourage customers to complete their purchases. This can increase revenue, reduce costs, and enhance customer satisfaction. Additionally, by reducing the number of incomplete purchases, businesses can reduce cart abandonment rates and increase customer loyalty. In summary, predictive models provide businesses with valuable insights into customer behavior that can help them make data-driven decisions and improve their bottom line.

6. Summary and Conclusions

In conclusion, our project demonstrates the effectiveness of machine learning models in predicting customer behavior in e-commerce. By analyzing customer behavior data, we were able to identify patterns that can help businesses understand what drives customers to complete their purchases. Our analysis found that logistic regression outperformed both random forest and neural network models in predicting customer behavior. Additionally, we identified the activity count feature as the most important variable in predicting whether a customer will make a purchase after adding a product to their cart.

The business value of accurately predicting customer behavior is significant, and businesses in the e-commerce industry can use predictive models to optimize their sales strategies and offer personalized incentives to encourage customers to complete their purchases. Our project provides a framework for businesses to develop predictive models that can enhance their understanding of customer behavior and increase their revenue and efficiency.

Reference :

1. Data source :
    - https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store
    - https://rees46.com/
2. Data source :
https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/