Analysing eCommerce behavior

From data exploration to modeling

Understanding eCommerce Behavior

Why this analysis is important

- eCommerce is a rapidly growing industry.
- Companies seek insights into user behavior for better decision-making.
- Competitive advantage through data-driven strategies.

What we aim to achieve

- Predict user purchases accurately.
- Understand user interactions and preferences.
- Identify actionable insights for business growth.

Who benefits from our analysis

- eCommerce companies and retailers.
- Marketing and sales teams.
- Data analysts and scientists.
- End-users for improved user experience.

Limitations of our study

- Data availability and quality.
- Resource and time constraints.

Problem formulation

Problem Type

- Binary Classification: Predicting Purchases (1) or No Purchases (0).

Objective

- Develop a model to understand and predict user purchase behavior based on user interactions.

Method

- Utilize supervised machine learning algorithms for predictive analysis.

Dataset description

Dataset Overview

- Kaggle Dataset: eCommerce Behavior Data from a Multi-Category Store.
- Contains extensive user behavior data.
- Features include event type, time, product details, and more.

Data Size : 67501979 rows, 9 columns.

Data Types: Numeric, Categorical, Datetime.

Data Sources: Acquired from an eCommerce platform.

Data Challenges

- Data preprocessing required (e.g., handling missing values, encoding categorical data).

Preparing data for analysis

Data Wrangling Steps

Data Loading
- Importing the dataset using Pandas.

Handling Missing Data
- Identifying and dealing with missing values.

Data Cleaning
- Removing duplicates.

Feature Engineering
- Creating new features for analysis (e.g., purchases per session, views per session).

Encoding Categorical Data
- Label encoding for categorical variables.

Standardization
- Scaling numerical features for consistency.

Data Transformation
- Converting datetime features.

Key Takeaways

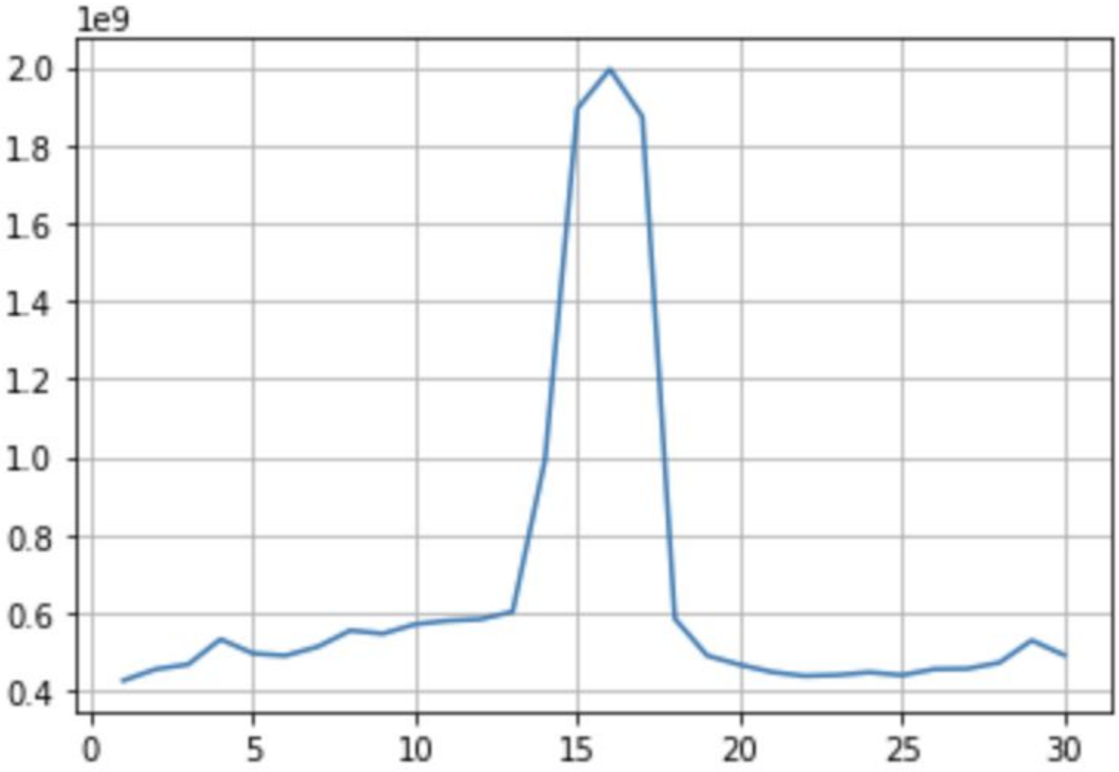
- Cleaned and prepared dataset for analysis.
- Engineered features to enhance predictive power.
- Ensured data consistency through encoding and scaling.
- Transformed datetime features for analysis.

Exploratory analysis

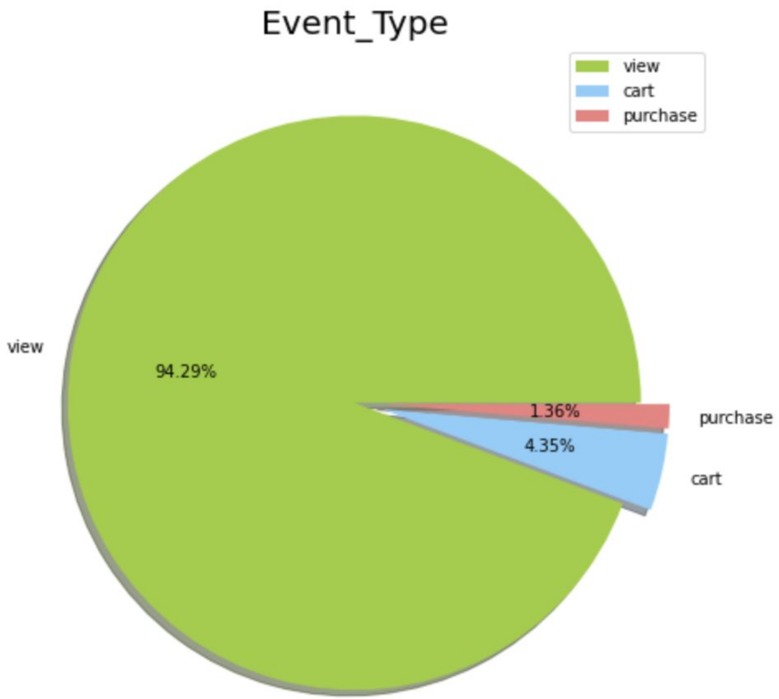
Sales trend:

purchases ($) per day of

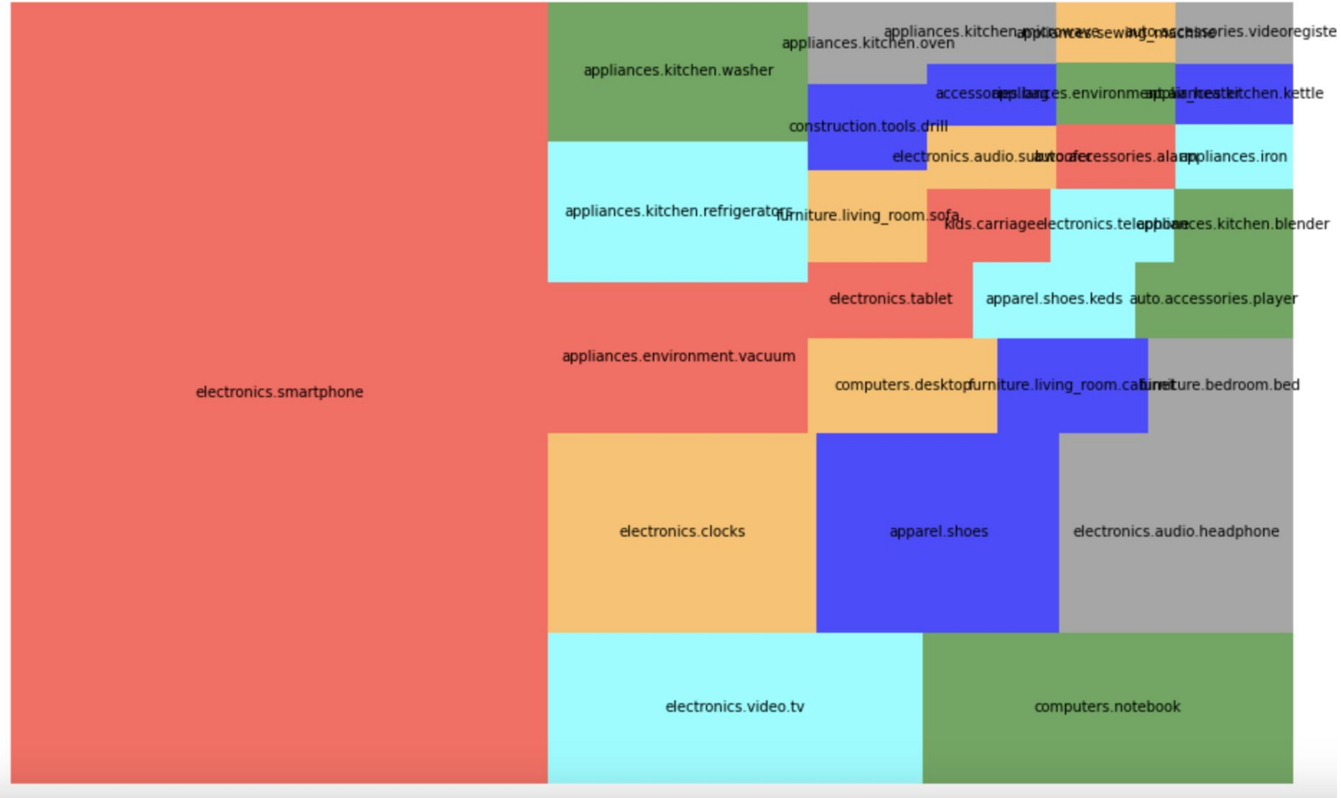
the month



User activities

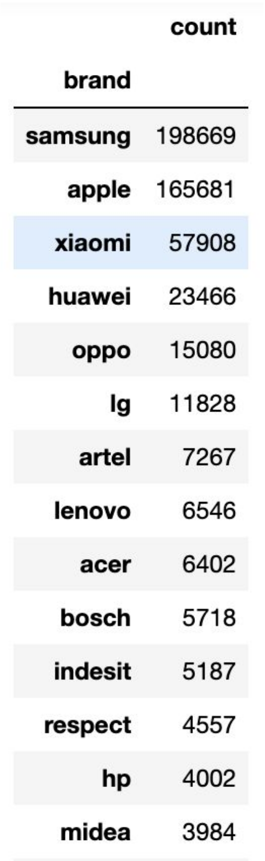


Product Categories



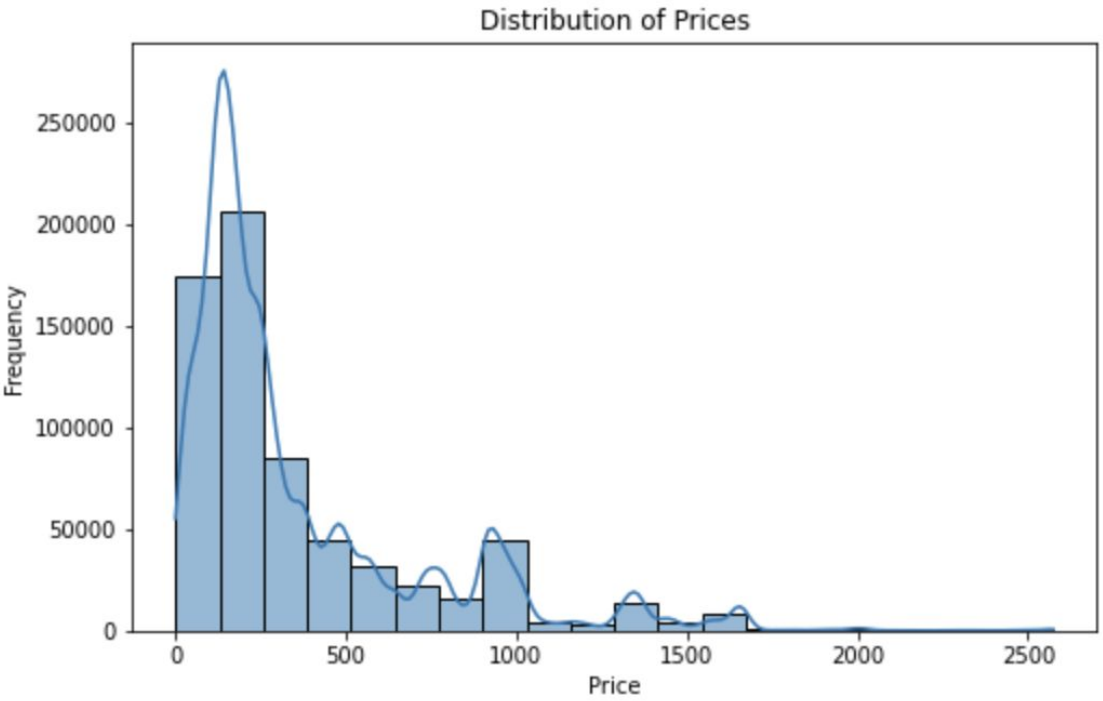
Brand preferences

Quantity of items of the most popular brands



	count
brand	
samsung	198669
apple	165681
xiaomi	57908
huawei	23466
oppo	15080
lg	11828
artel	7267
lenovo	6546
acer	6402
bosch	5718
indesit	5187
respect	4557
hp	4002
midea	3984

Price Distribution



Modeling process

Data Preprocessing

- Data scaling and transformation.
- Handling categorical variables.

Model Selection

- Choosing appropriate algorithms (XGBoost, Logistic regression).

Model Training

- Splitting data into training and testing sets.
- Training models on the dataset.

Performance Metrics

- Metrics used to evaluate model performance (Accuracy).

Reasons to choose XGBoost:

Handling Imbalanced Data

Accuracy

Feature Importance

Speed and Efficiency

Interpretability

Reasons to choose Logistic regression:

Interpretability

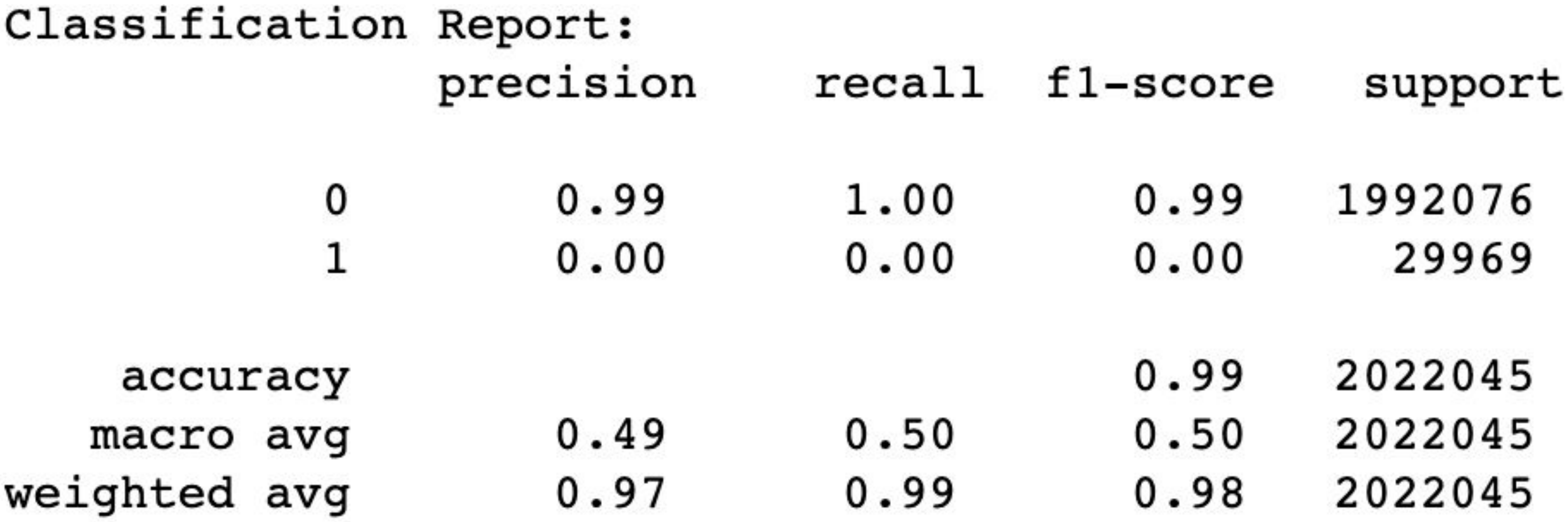
Simplicity

Efficiency

Low Variance

Well-Suited for Binary Classification

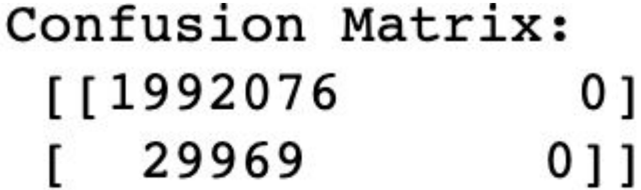
Performance Metrics



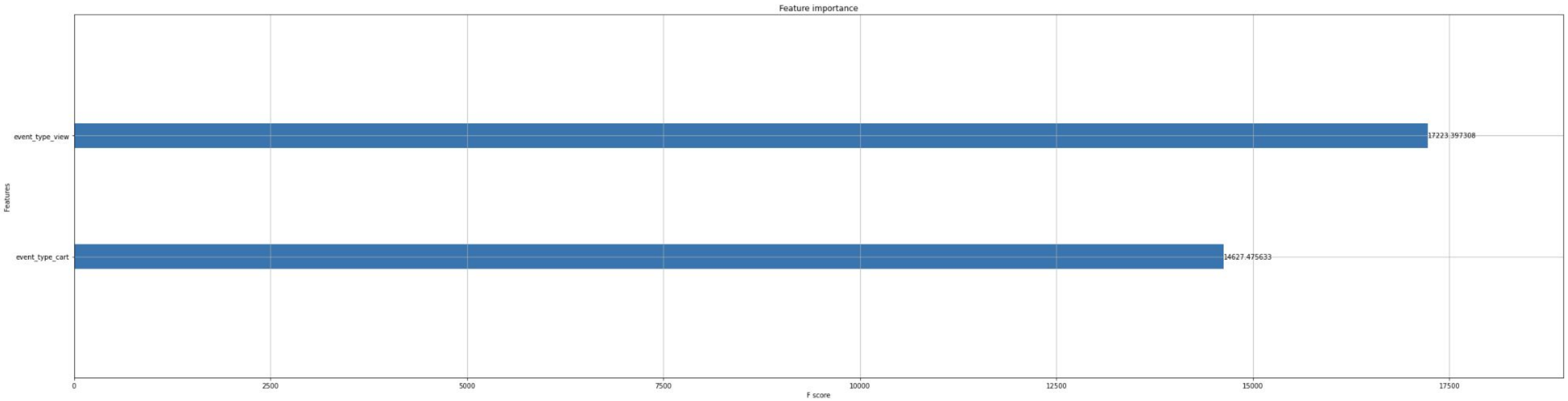
```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99   1992076
           1       0.00      0.00      0.00     29969

    accuracy                           0.99   2022045
   macro avg       0.49      0.50      0.50   2022045
weighted avg       0.97      0.99      0.98   2022045
```



```
Confusion Matrix:
 [[1992076       0]
 [  29969       0]]
```



“View” and “Cart” features strongly influenced purchase decisions.

Practical considerations and future work

Scalability: Ensure that systems can handle larger datasets and increased computational demands as the business grows. Probably need to implement distributed computing, or data partitioning.

Real-time Analytics: If real-time insights are critical, invest in streaming data processing and real-time analytics tools. This allows for immediate responses to changing user behavior and market trends.

Cost Management: Keep an eye on the cost of maintaining and running machine learning models. Optimize cloud resources and consider cost-effective alternatives for storing and processing data.

User Experience (UX): Prioritize the user experience when implementing data-driven recommendations or marketing strategies. Ensure that recommendations are relevant and enhance the overall user journey.

Suggestions For Improvement

User Segmentation:

- Explore user segmentation techniques to tailor marketing strategies more effectively. Identifying distinct user groups and customizing marketing approaches can lead to higher conversion rates.

Advanced Analytics:

- Investigate advanced analytics methods such as deep learning or time series forecasting, especially if the dataset grows in complexity or if the business needs evolve.

Data Augmentation:

- If more data is obtainable, consider data augmentation techniques to expand the dataset. Augmenting data can help the model generalize better and improve its predictive power.