Engage2Value : ML Prediction Pipeline

By Piush Das

Project Overview

This project aims to predict the purchaseValue of customers based on their multi-session behavior across digital touchpoints on a large-scale e-commerce platform. The notebook follows a complete end-to-end machine learning workflow including data loading, exploratory data analysis (EDA), categorical and numerical feature analysis, outlier study, preprocessing, model building, hyperparameter tuning, and model comparison.

The primary objective, as defined in the competition, is to model session-level behavioral patterns such as engagement metrics, traffic sources, device characteristics, and geographical indicators to estimate the purchase potential of each session. The problem is formulated as a supervised regression task and evaluated using the R^2 score between predicted and true purchaseValue.

The notebook is structured systematically to reflect the Machine Learning Development Life Cycle (MDLC), ensuring reproducibility, interpretability, and robust model performance on a high-dimensional, real-world clickstream dataset.

Dataset Description

The dataset captures detailed session-level information from a digital commerce platform where each row corresponds to a unique user session. It includes anonymized user interactions such as browser type, traffic acquisition channels, device specifications, engagement metrics, and geographical attributes.

The dataset provided consists of:

- train.csv (training data with purchaseValue)
- test.csv (test data without target)
- sample_submission.csv (submission format)

Key feature groups in the dataset include:

User Behavior & Session Metrics

- totalHits, pageViews, totals.bounces, new_visits, totals.visits
- sessionNumber and sessionStart
  These features represent user engagement intensity and session activity patterns.

Device & Technical Attributes

- deviceType, os, browser, screenSize, device.browserSize, device.language
- browserMajor and other device-level descriptors
  These capture the technical environment of the user which can influence browsing and purchasing behavior.

Traffic & Marketing Source

- userChannel, trafficSource, trafficSource.medium
- trafficSource.keyword, campaign, referralPath
- gclIdPresent and ad-related attributes
  These features provide insights into how users arrived at the platform and their acquisition journey.

Geographical Context

- geoNetwork.city, locationCountry, continent, region, metro
- geoCluster and locationZone
  These variables help identify regional purchasing trends and behavioral segmentation.

Identifiers

- userId and sessionId for multi-session tracking and behavioral continuity.

Target Variable

- purchaseValue: The total monetary amount spent during a session, which is the variable to be predicted.

The dataset is highly complex due to:

- 37 categorical columns (object type)
- High missing values in multiple traffic-related features
- Skewed numerical distributions
- Highly imbalanced target with many zero-purchase sessions

Loading The Dataset

The project begins with loading the provided competition datasets: train.csv, test.csv, and sample_submission.csv, which contain session-level behavioral data from a digital commerce platform. Each row represents a unique user session with multiple features describing engagement, device characteristics, traffic sources, and geographical context.

The primary objective defined in the problem statement is to predict the **purchaseValue**, which represents the total amount spent by a user during a session. This establishes the task as a supervised regression problem evaluated using the R^2 score.

During initial dataset loading and inspection, the structure of the training dataset revealed a high-dimensional tabular dataset containing a mix of numerical, categorical (object), and boolean features. The dataset is complex due to anonymized fields, high cardinality categorical columns, and significant missing values across multiple traffic and advertisement-related features.

Importing Necessary Libraries

The notebook imports essential Python libraries required for data analysis, preprocessing, and model development. These include libraries for:

- Data manipulation and analysis (pandas, numpy)
- Visualization and EDA (matplotlib, seaborn)
- Machine Learning pipeline construction (scikit-learn)
- Model building and evaluation (ensemble and regression models)
- Preprocessing tools such as encoders, scalers, and transformers

The use of a structured ML pipeline with preprocessing utilities indicates a well-designed workflow that minimizes data leakage and ensures reproducibility of results.

Exploratory Data Analysis (EDA)

Extensive Exploratory Data Analysis (EDA) was conducted to understand the dataset's structure, feature distributions, and data quality issues.

From the notebook analysis:

There are four kinds of data types present in the training set:

- Object: 37 columns
- Integer: 9 columns
- Float: 5 columns
- Boolean: 1 column

This clearly shows that the dataset is heavily dominated by categorical (object) features, which increases preprocessing complexity and requires careful encoding strategies.

A critical observation from the EDA is the presence of severe missing values in several columns. Specifically, columns such as:

- trafficSource.adContent
- trafficSource.adwordsClickInfo.isVideoAd
- trafficSource.adwordsClickInfo.page
- trafficSource.adwordsClickInfo.adNetworkType
- trafficSource.adwordsClickInfo.slot

contain more than **90% missing values**, making them largely uninformative and potentially harmful for model performance if not handled properly.

Additionally, the target variable purchaseValue shows a highly skewed distribution with a large number of zero-value sessions, which is typical in e-commerce datasets where only a small fraction of sessions result in purchases.

Analysis of Categorical Column

A detailed categorical feature analysis was performed in the notebook to understand behavioral and platform patterns.

trafficSource.isTrueDirect

- This column contains approximately **63% null values**
- For the non-null values, the dominant value is only **True**
- This indicates extremely low variability and weak predictive usefulness

browser

- **Chrome** is the most used browser among 34 unique browser values
- Followed by Safari and Firefox
- This shows strong platform concentration and suggests that browser type may indirectly influence engagement and purchase behavior

device.screenResolution & ScreenSize

These features provide insight into user device characteristics and user interface interaction patterns, which can influence browsing behavior and purchase decisions.

Geo Cluster

- The dataset is uniformly sampled across all five regions
- Each region contributes approximately 20% of the total data

  This indicates balanced geographical representation and reduces regional bias in model training.

geoNetwork.networkDomain

- The distribution is nearly uniform across domain1, domain2, and domain3
- Each contributes roughly one-third of the total observations

  This suggests no single domain dominates the traffic distribution.

device.mobileDeviceBranding & marketing-related features

These features help capture device-level behavioral patterns and user segmentation across different marketing channels and hardware environments.

Overall, categorical analysis highlights that:

- High cardinality features exist
- Some columns have low variance
- Some columns have extreme missing values
- Certain features like browser and traffic source show strong dominance patterns

Analysis of Numerical Column

The numerical feature analysis focused on session engagement metrics and behavioral indicators such as:

- totalHits
- pageViews
- sessionNumber
- totals.visits
- new_visits

These numerical columns reflect user interaction intensity and session activity.

Key observations:

- Engagement features such as totalHits and pageViews exhibit skewed distributions
- Most sessions have low engagement, while a small number show extremely high activity
- This heavy-tailed distribution aligns with real-world clickstream data
- Higher engagement sessions are more likely to result in non-zero purchaseValue

The skewness in numerical features indicates the need for scaling and potential transformation to stabilize model learning.

Studying Outliers

Outlier analysis was conducted to identify extreme behavioral patterns in engagement-related numerical columns.

Findings from the notebook:

- totalHits and pageViews contain significant outliers
- A small number of sessions show extremely high values compared to the majority
- These outliers likely correspond to highly engaged users or bot-like sessions

Given the business context, these outliers are not removed blindly because:

- High engagement can strongly correlate with high purchaseValue
- Removing them may reduce model predictive power

Instead, the preprocessing pipeline is designed to handle skewness rather than aggressively removing outliers.

PreProcessing

A structured preprocessing pipeline was implemented in the notebook to handle the complex dataset.

Key preprocessing steps include:

Handling Missing Values

- Columns with more than 90% missing values were identified as low-information features
- Null-heavy advertisement columns were treated carefully or excluded
- Remaining missing values were handled using appropriate imputation strategies

Categorical Encoding

- The dataset contains 37 object-type columns
- Encoding techniques were applied to convert categorical variables into machine-readable format
- High-cardinality categorical features were processed using efficient encoding methods

Feature Transformation

- Numerical features were scaled using appropriate scaling techniques
- ColumnTransformer was used to apply separate transformations to numerical and categorical columns
- This ensured a clean and reproducible preprocessing workflow

Data Leakage Prevention

The preprocessing pipeline was integrated into the model pipeline, ensuring that transformations were learned only from training data and applied consistently to validation and test sets.

Model Building

The notebook implements regression-based machine learning models to predict purchaseValue.

Key modeling characteristics:

- The problem is treated as a supervised regression task
- The evaluation metric used is **R^2 score**, as specified in the competition
- Pipeline-based modeling approach was used (Preprocessing + Model)

Tree-based ensemble models were selected because:

- They handle mixed data types well
- They capture non-linear relationships
- They are robust to outliers and skewed distributions
- They perform strongly on high-dimensional tabular datasets

HyperParameter Tuning

Hyperparameter tuning was performed to improve model generalization and predictive accuracy.

Key tuning considerations:

- Number of estimators
- Maximum tree depth
- Learning rate
- Subsampling parameters
- Regularization parameters

The tuning process focused on maximizing the R² score while preventing overfitting, especially considering the skewed target distribution and high-dimensional feature space.

Model Comparison

Multiple models were compared based on their R² performance and generalization capability on validation data.

Key comparison insights:

- Ensemble tree-based models outperformed basic baseline models
- Proper preprocessing significantly improved model performance
- Models capturing non-linear relationships between engagement, traffic source, and device features performed better
- Feature engineering and missing value handling had a direct impact on prediction accuracy

Insights

1. The target variable purchaseValue is highly skewed with a large number of zero-purchase sessions, making this a challenging regression problem.
2. Engagement metrics such as totalHits and pageViews are strong indicators of purchase behavior.
3. Chrome dominates browser usage, indicating platform concentration in user traffic.
4. Several advertising-related columns contain more than 90% missing values and contribute minimal predictive value.
5. High-cardinality categorical features significantly influence model complexity and preprocessing design.
6. Uniform geo cluster distribution indicates balanced regional representation in the dataset.
7. Outliers in engagement metrics represent highly active users who may contribute significantly to revenue prediction.

Recommendation

1. Drop or carefully treat columns with extremely high missing values (>90%) to reduce noise and model complexity.
2. Apply advanced encoding techniques for high-cardinality categorical features to improve model learning.
3. Consider log transformation of purchaseValue to handle heavy skewness and improve regression stability.
4. Focus marketing strategies on high-engagement users (high pageViews and totalHits) as they show higher purchase potential.
5. Use ensemble models such as Gradient Boosting, XGBoost, or LightGBM for better performance on tabular behavioral data.
6. Implement feature importance analysis to identify the most revenue-driving behavioral and traffic features.

Conclusion

This project successfully developed a comprehensive machine learning pipeline to predict customer purchaseValue using multi-session behavioral data from a digital commerce platform. The notebook followed a structured workflow including EDA, categorical and numerical analysis, outlier study, preprocessing, model building, hyperparameter tuning, and model comparison aligned with the competition objective of maximizing R^2 score.

The analysis revealed that user engagement metrics, traffic sources, and device attributes play a critical role in predicting purchase behavior, while several high-missing advertisement features provide limited value. The use of a robust preprocessing pipeline and ensemble modeling approach enabled effective handling of high-dimensional categorical data, skewed numerical distributions, and outliers.

Overall, the model demonstrates strong potential for real-world applications such as customer value prediction, marketing optimization, revenue forecasting, and personalized engagement strategies in digital commerce environments.