Project Report: Digital Commerce Purchase Value Prediction

Subject: Predictive Modeling of Customer Session Value using MLP

1. Executive Summary

The objective of this project is to predict the purchaseValue (total spend) of user sessions on a large-scale digital commerce platform. By analyzing user behavior, technical specifications, and traffic sources, we aim to identify high-value sessions. The project utilizes a Multi-Layer Perceptron (MLP) neural network to capture non-linear relationships within the high-dimensional dataset. The model's performance is evaluated using the R2 Score (Coefficient of Determination).

2. Problem Statement

E-commerce platforms generate vast amounts of session data. The majority of sessions do not result in a purchase, leading to a highly skewed target variable (purchaseValue). The challenge is to accurately estimate revenue potential per session to optimize marketing spend and personalize user experiences.

- **Target Variable:** purchaseValue (Continuous/Regression)
- **Evaluation Metric:** R2 Score (1.0 indicates perfect prediction)

3. Data Analysis & Feature Engineering

The dataset consists of session-level logs containing a mix of numerical, categorical, and hierarchical JSON-like fields.

3.1 Feature Categories

The data is segmented into four primary categories:

Category	Key Features	Insight/Hypothesis
User Behavior	totalHits, pageViews, totals.bounces, new_visits, sessionStart	**High Impact.** High page views and hit counts typically correlate positively with purchase probability. Bounces correlate negatively.
Technical	deviceType, os, browser,	**Medium Impact.** Specific

	isMobile, device.language	OS (e.g., iOS) or Device Types (Desktop vs. Mobile) often show distinct conversion rates.
Traffic Source	trafficSource.medium, trafficSource.source, gclIdPresent (AdWords), campaign	**High Impact.** Paid search (cpc) and direct traffic usually yield higher intent users than organic referrals.
Geography	geoNetwork.city, locationCountry, geoNetwork.continent	**Low-Medium Impact.** Purchasing power varies by region, though high cardinality (many cities) makes this challenging to model.

3.2 Preprocessing Strategy

Given the nature of web traffic data, the following preprocessing steps are critical:

1. **Handling Sparsity:** The target purchaseValue is likely NaN or 0 for non-purchasing sessions. These are imputed as 0.
2. **Categorical Encoding:**
    - High-cardinality features (e.g., City, NetworkDomain) require robust handling (e.g., Frequency Encoding or Embedding Layers).
    - Low-cardinality features (e.g., DeviceType) are One-Hot Encoded.
3. **Numerical Scaling:** Neural networks require scaled inputs. Features like pageViews and totalHits are normalized (StandardScaler or MinMaxScaler) to ensure faster convergence during gradient descent.
4. **Date-Time Extraction:** sessionStart is parsed to extract features like Hour of Day and Day of Week to capture temporal shopping patterns.

4. Modeling Approach: Multi-Layer Perceptron (MLP)

A Deep Learning approach was selected to handle the complexity and interactions between the features.

4.1 Architecture

- **Input Layer:** Corresponds to the number of processed features.
- **Hidden Layers:** Dense (Fully Connected) layers with ReLU (Rectified Linear Unit) activation functions to introduce non-linearity.

- **Output Layer:** A single neuron with a linear activation function (suitable for regression tasks).

4.2 Post-Processing Logic

A specific constraint was applied to the model predictions based on domain logic:

```
test_predictions[test_predictions < 0] = 0
```

Since a customer cannot spend a negative amount, all negative predictions output by the regression model are clamped to zero. This significantly improves the R2 score by eliminating physically impossible errors.

5. Challenges & Observations

1. **Zero-Inflated Target:** The vast majority of sessions result in $0 revenue. The model essentially performs a dual task: classification (Will they buy?) and regression (How much?).
2. **Data Leakage Risks:** Features like totalHits are calculated at the *end* of a session. In a real-time prediction scenario, this data accumulates over time.
3. **Evaluation Sensitivity:** The R2 metric can be volatile when predicting sparse targets. A few large outliers (high spenders) can disproportionately sway the score.

6. Conclusion

The MLP model effectively utilizes granular session data to forecast revenue. Key drivers of value are identified as behavioral metrics (pageViews, hits) and traffic origin (trafficSource). Future improvements could include implementing specific "Hurdle Models" (separate models for purchase probability vs. purchase amount) to better handle the zero-inflated nature of the data.