Early Customer Signals & Long-Term Value

An end-to-end e-commerce analysis using SQL, Python, and Tableau

Problem Statement

Can we identify high-value customers early in their lifecycle using only initial behavioural signals?

Many businesses wait months before segmenting customers by lifetime value.
The objective here was to test whether **early activity (first 30 days)** is statistically informative enough to identify long-term value segments.

Context from Previous Models

Model 1 showed that:

- Discounts do not increase order-level revenue.
- Price and quantity are far stronger revenue drivers.

Model 2 showed that:

- Discount exposure does not necessarily increase repeat behaviour.
- Medium-discount customers demonstrated higher engagement but not necessarily higher profitability.

This raised a new question:
Instead of focusing on discounts, can early behaviour signal long-term customer value more effectively?

Data & Feature Engineering

Built from the existing SQLite schema and customer summary table.
For each customer, early lifecycle features were engineered:

- Orders in first 30 days
- Revenue in first 30 days
- Average discount received in first 30 days
- Income band
- Gender
- Discount bucket
- Total long-term revenue (target outcome)

A “high-value customer” was defined as customers above the 75th percentile of total revenue.

High-value cutoff: ₹40,700 (approx.)

Class distribution:

- 378 non-high-value customers
- 126 high-value customers

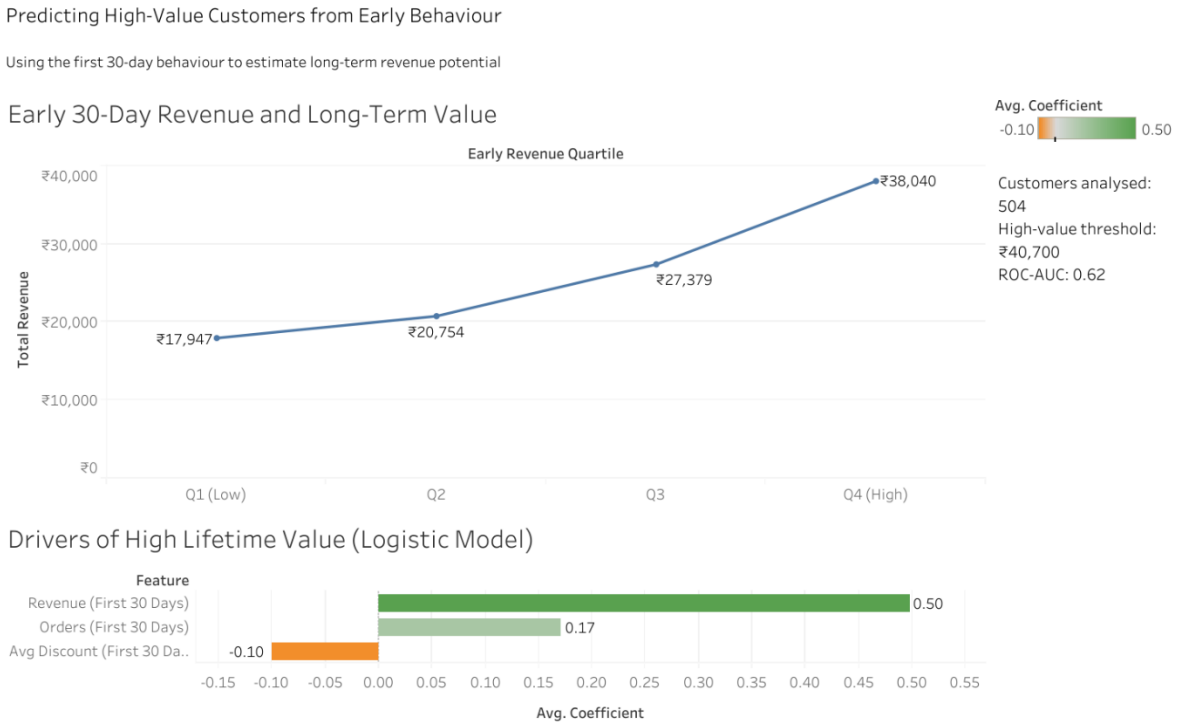
Statistical Approach

A logistic regression model was used to test whether early signals predict high-value classification.

This was treated as a statistical test of predictive strength, not as a black-box ML exercise.

Three core predictors were evaluated:

- Revenue in first 30 days
- Orders in first 30 days
- Average discount in first 30 days



Link: https://tinyurl.com/early-behaviour-dashboard

Model Performance

Test Set Metrics:

- Accuracy: 82%
- ROC-AUC: 0.62

However:

- Recall for high-value customers was low.
- Model performs better at identifying non-high-value customers.
- Predictive strength is moderate, not strong.

Interpretation:

Early signals do contain information, but they are not sufficient alone for robust classification.

Coefficient Interpretation
Standardized Logistic Coefficients:

- Revenue (first 30 days): +0.497
- Orders (first 30 days): +0.170
- Average Discount (first 30 days): -0.100

Implications:

- Early revenue is the strongest signal of long-term value.
- Order frequency has a smaller but positive effect.
- Higher early discounts slightly decrease the probability of long-term high value.

This reinforces findings from Models 1 and 2:
Discounting does not appear to create structurally strong customers.

Strategic Insight
Value appears to be driven by:

- Early willingness to pay
- Initial purchase intensity
- Intrinsic customer behaviour

Rather than discount exposure.

This suggest businesses should:

- Identify strong early spenders
- Prioritize premium engagement for them
- Avoid over-subsidizing low-margin segments via discounts

Limitations

- Dataset is synthetic but structured.
- Class imbalance limits high-value recall.
- Only 30-day signals were used.
- No behavioural sequence modelling.
- No cross-validation tuning applied.

Business Takeaway
While early revenue is directionally predictive of long-term value, discounts do not meaningfully create high-value customers.

Targeting should prioritize identifying naturally high-intent customers rather than stimulating artificial demand through discount depth.