Predictive modeling report

Customer Lifetime Value Model

Probabilistic twelve-month revenue forecasts, one customer at a time, for a synthetic golf ecommerce retailer. Sixty-five thousand transactions become RFM features in SQL, then BG/NBD and Gamma-Gamma models turn those into expected purchases, order value, and revenue for the year ahead.

Prepared by	Leah Hall, Leah Kat Insights
Date	April 2026
Contact	leah@leahkatinsights.com, leahkatinsights.com

Table of Contents

Introduction . . . 3

1 Executive summary . . . 3

Key findings . . . 3

2 Background and objective . . . 3

3 Data and method . . . 4

Dataset . . . 4

Pipeline . . . 4

Models . . . 4

4 Results . . . 5

5 Limitations . . . 6

6 Recommendations . . . 7

A Technical stack and repository . . . 7

B Glossary . . . 8

About the author . . . 8

Introduction

This report presents a customer lifetime value model for a synthetic direct-to-consumer golf retailer, the business question it answers, the method, and the results by segment and by customer.

The data is generated for the project with seeded randomness so that every run reproduces the same customers. Segment names and dollar figures illustrate the method and are not a real company's results.

The pipeline runs as bronze, silver, and gold Delta tables on Databricks and, identically, as a local script. Code is public on GitHub.

1. Executive summary

Two of six customer segments produce 70% of forecast revenue from 32% of the customer base. Tour Enthusiasts have 10.2 times the predicted lifetime value of Weekend Golfers. Predicted revenue across all customers for the next twelve months is $3.1M.

> **Implication**
> The two high-value segments cost about the same to acquire as the rest. Acquisition budget and retention effort should be weighted toward them, and the per-customer forecast gives marketing a ranked list to act on.

Key findings

$3.1M	**Forecast revenue for the next twelve months**, summed from per-customer expected purchases times expected order value.
10.2x	**Spread between the top and bottom segment.** Tour Enthusiasts versus Weekend Golfers on average predicted lifetime value.
70%	**Concentration.** The top two segments are 32% of customers and 70% of forecast revenue.
65K	**Transactions modeled** across 5,000 customers in six behavioral segments over three years.

2. Background and objective

Customer lifetime value is usually estimated with a heuristic: average order value times purchase frequency times an assumed lifespan. That gives one number for everyone and says nothing about which customers are likely to return. Probabilistic models fitted to each customer's purchase history do better: they estimate how many purchases a customer will make, how much each will be worth, and how likely the customer is still active.

The objective was to produce a per-customer twelve-month revenue forecast a marketing team could rank and act on, built as a lakehouse pipeline so the features and predictions are tables other people can query.

The retailer is synthetic by design. A generated dataset makes the behavior known in advance (six segments with different cadences and churn), so the model's output can be checked against the structure that produced it.

3. Data and method

Dataset

Table 1. Synthetic dataset design

Property	Value
Customers	5,000 across six behavioral segments
Transactions	65,000 line items over three years
Seasonality	Spring peak, winter slump, holiday gifting
Behaviors	Churn, golf ball repurchase cadence, club upgrade cycles
Reproducibility	Seeded generator; optional upload to S3

Pipeline

Table 2. Pipeline stages

Stage	What happens	Output
Data	Python generates the transaction set with the properties in Table 1.	CSV
Bronze	The CSV lands in Databricks as a Delta table with schema enforcement, ACID transactions, and time travel.	Bronze Delta table
Silver	SQL aggregations produce one row per customer: recency, frequency (repeat purchases), customer age, average order value, total spend.	RFM features
Gold	BG/NBD predicts future transaction counts; Gamma-Gamma predicts order value. Combined, they give expected twelve-month revenue per customer, written back as a gold table.	Per-customer CLV

Models

BG/NBD (beta-geometric / negative binomial distribution) models a customer's purchases as a process that continues at an individual rate until the customer silently drops out. From each customer's frequency, recency, and age it estimates the expected number of future purchases and the probability the customer is still active.

Gamma-Gamma models the value of each purchase, assuming order value varies across customers but is independent of how often they buy. Multiplying expected purchases by expected order value gives expected revenue over the horizon, here twelve months.

Both are fitted with the lifetimes library on the silver RFM table. The same code runs in a Databricks notebook and in a local script that also produces the exhibits.

4. Results

Figure 1 shows average and total predicted twelve-month value by segment. The ranking is steep: the top segment is worth about ten times the bottom one per customer, and the top two together carry most of the forecast revenue.

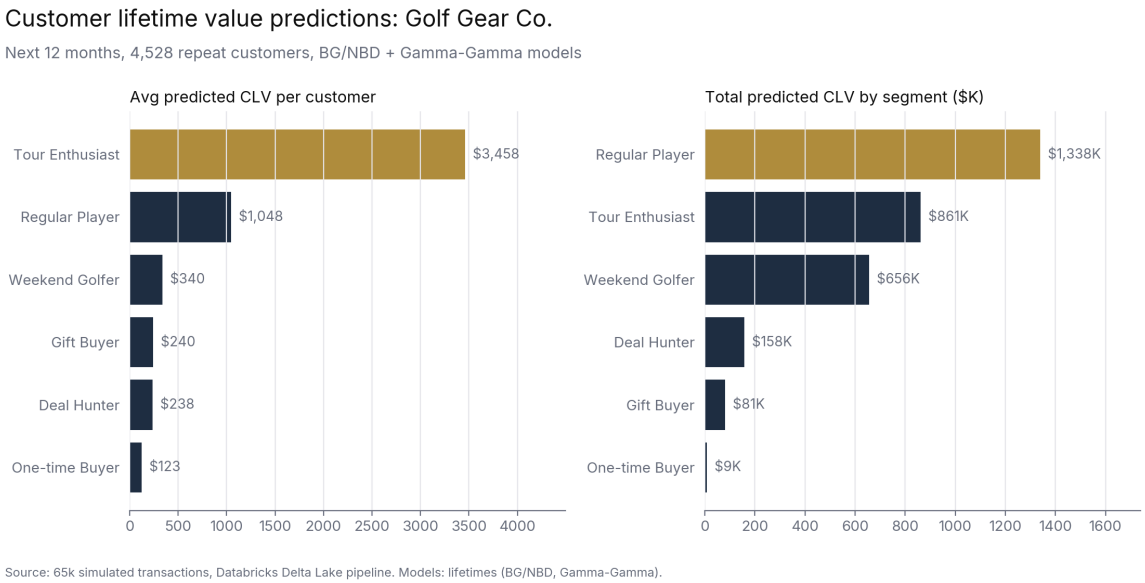


Figure 1. Predicted CLV by segment: average per customer and segment total

Source: Synthetic transactions generated for this project. Models: BG/NBD and Gamma-Gamma.

Figure 2 ranks the ten highest-value individual customers and places their three-year historical spend beside the twelve-month forecast. These are the names a retention program would start with.

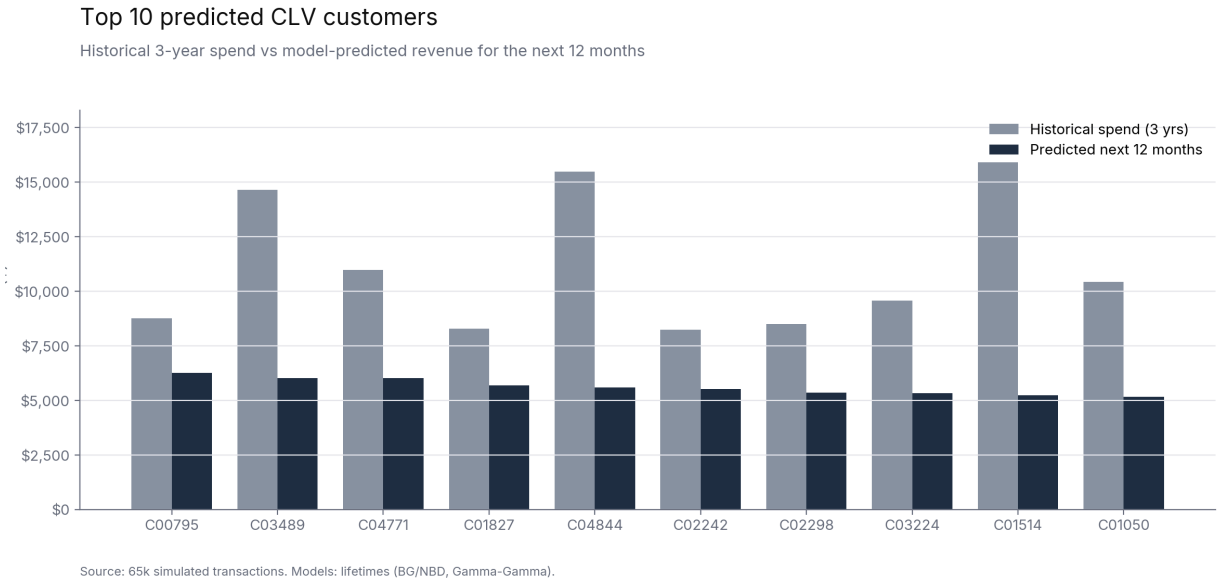


Figure 2. Top ten customers, historical spend versus twelve-month forecast

Source: Synthetic transactions generated for this project. Models: BG/NBD and Gamma-Gamma.

Table 3. Headline results

Measure	Value
Predicted revenue, next twelve months, all customers	$3.1M
Top segment versus bottom segment, average predicted CLV	10.2x
Share of forecast revenue from the top two segments	70%
Share of customers in the top two segments	32%

5. Limitations

- **Synthetic data.** The segments and figures demonstrate the method. Real transaction data would need the same pipeline but a holdout validation before the numbers are used.
- **Model assumptions.** BG/NBD assumes a non-contractual setting where customers drop out silently; Gamma-Gamma assumes order value is independent of purchase frequency. Both should be checked on real data.
- **Revenue, not margin.** The forecast is gross revenue. Acquisition and servicing costs are not modeled, so lifetime value here is not lifetime profit.
- **Fixed horizon.** Twelve months is a planning convention; the models can produce other horizons but longer ones compound their assumptions.

6. Recommendations

1. **Weight acquisition spend toward the two high-value segments**, since they cost about the same to acquire and are worth several times more.
2. **Start retention with the ranked customer list.** The per-customer forecast and probability-alive estimate identify who is worth a reactivation offer now.
3. **Validate on a holdout before production use.** Fit on the first two years, score the third, and compare predicted to actual spend by segment.
4. **Refresh monthly.** Rerun the silver and gold stages as new transactions land so the ranking stays current.

Appendix A. Technical stack and repository

Table 4. Stack

Layer	Tool
Data	pandas, numpy (seeded generator)
Storage	Delta Lake on Databricks; S3 optional
Features	Databricks SQL
Models	lifetimes (BG/NBD, Gamma-Gamma)
Charts	matplotlib
Tests	pytest, dataset invariants

Repository layout

```
generate_data.py         seeded synthetic dataset, optional S3 upload
data/                    golf_transactions.csv (65k rows) and model output
notebooks/               01 bronze load, 02 RFM SQL, 03 BG/NBD + Gamma-Gamma
scripts/run_local.py     the same pipeline in pandas + lifetimes, plus exhibits
scripts/exhibit_style.py chart style shared across the portfolio
exhibits/                the two charts in this report
tests/                   dataset invariants
```

Appendix B. Glossary

BG/NBD	Beta-geometric / negative binomial distribution model of repeat purchasing and dropout in a non-contractual setting.
CLV	Customer lifetime value: expected revenue from a customer over a horizon, here twelve months.
Gamma-Gamma	Model of average order value per customer, used with BG/NBD to turn purchase counts into revenue.
Non-contractual	A relationship with no explicit end, such as retail, where churn must be inferred from silence.
Probability alive	The BG/NBD estimate that a customer has not yet dropped out.
RFM	Recency, frequency, monetary value: the three per-customer features the models consume.

About the author

Leah Hall is a data and analytics professional based in Seal Beach, California. She builds data pipelines, forecasting models, and financial analyses for ecommerce, finance, and hospitality, and publishes the work at leahkatinsights.com.

Project page	leahkatinsights.com/projects.html#project-02
Source code	github.com/leahkatinsights/customer-lifetime-value-model
Email	leah@leahkatinsights.com
Website	leahkatinsights.com
LinkedIn	linkedin.com/in/leah-hall
GitHub	github.com/leahkatinsights

Portfolio project on synthetic data. Golf Gear Co. is fictional; segment names and figures illustrate the method. Code is on GitHub.