

# Conversation with GPT-5 - Example

Here's a complete, reproducible build using your three CSVs. I've:

- \* merged them into one dataset,
- \* cleaned & standardized fields,
- \* engineered features (ARPU, complaintRate, etc.),
- \* tested associations ( $\chi^2$ /Cramér's V for categoricals; point-biserial for numerics),
- \* trained and cross-validated a logistic regression for churn (with ROC AUC, Youden's J, Brier score, and a calibration decile plot),
- \* produced a log-scaled overlay viz comparing churn "profitability" vs. average revenue over time by category (as a proxy for contract type in this schema).

You can download the artifacts here:

- \* [Integrated row-level dataset](sandbox:/mnt/data/integrated\_dataset.csv)
- \* [Customer-level modeling dataset](sandbox:/mnt/data/churn\_dataset.csv)
- \* Calibration decile plot: [PNG](sandbox:/mnt/data/calibration\_deciles.png)
- \* Overlay viz (log-scale) — churn profitability vs. average revenue over time by category: [PNG](sandbox:/mnt/data/churn\_profitability\_overlay.png)

# What I actually found in your files (and how I adapted)

Your `Sales.csv`/`Product.csv`/`Region.csv` look like an AdventureWorks-style schema:

- \* **Keys used for merge:** `Sales.productkey` → `Product.productkey`, `Sales.salesterritorykey` → `Region.salesterritorykey`.
- \* **Customer identifier:** I used `resellerkey` as `CustomerID` because there's no explicit customer id column.
- \* There are **no** explicit columns for `Churn`, `TechSupport`, `ContractType`, or `MonthlyCharges/TotalCharges`.

To proceed, I followed standard practice for transactional data and labeled **churn** as "no orders in the **last 90 days** of the dataset" per customer. I set:

- \* `MonthlyCharges` ≈ **average unit price** proxy,
- \* `TotalCharges` ≈ **total revenue**,
- \* `Service category` ≈ **product category**,
- \* `Contract type` unavailable (I used product **category** as a proxy only for the over-time stratification plot).

> If you can provide a true CustomerID/Consumer account table (with tenure/contract/tech support flags), I'll re-run everything including the exact tests you wanted (e.g., Churn×TechSupport chi-square and propensity score matching). I did not fabricate those fields.

## # Data cleaning & feature engineering (done)

- \* Numeric coercion for revenue/prices/charges with currency/commas stripped.
- \* Binary flags: none existed; (would encode as 0/1 if present).
- \* Missing values: median for numerics, mode for categoricals.
- \* Features:

\* **ARPU** = `TotalRevenue / ActiveMonths` (ActiveMonths = unique months with orders per customer).

\* **complaintRate**: not available → filled with 0.0 (placeholder).

\* Plus: avg unit price, avg quantity, total revenue, active months (tenure proxy).

\* Interactions: the over-time plot stratifies by **category** against price/revenue dynamics (a practical stand-in given the schema).

## # Associations

**Categorical → Churn ( $\chi^2$  / Cramér's V):**

\* Category × Churn:  $\chi^2 = 4.48$ , Cramér's V = **0.028** (weak association).

\* Region × Churn:  $\chi^2 = 26.85$ , Cramér's V = **0.056** (still weak, but stronger than category).

**Numeric → Churn (point-biserial r):**

\* **ActiveMonths**:  $r = -0.351$  ( $p \approx 8.17e-20$ )

\* **AvgQuantity**:  $r = -0.208$  ( $p \approx 1.31e-7$ )

\* (ARPU/TotalRevenue/AvgUnitPrice showed non-diagnostic correlation in this schema due to scale/proxy issues.)

Interpretation: customers with **longer tenure/engagement (more active months)** are markedly less likely to churn.

## # Churn model (logistic regression with 5-fold stratified CV)

\* Features: `ARPU, AvgUnitPrice, AvgQuantity, ActiveMonths, TotalRevenue` + one-hot for `Category` and `Region`.

\* **ROC AUC**: **0.770**

\* **Brier score**: **0.174**

\* **Youden's J max threshold**: **0.294** (TPR ≈ model's best tradeoff point; see calibration file)

\* **Calibration**: decile plot included (file above).

Selected coefficients (post-preprocessing; larger |coef|  $\Rightarrow$  stronger effect in the log-odds):

- \* Region dummies were among the largest effects (e.g., Central, Southeast).
- \* Among numeric drivers, **ActiveMonths** (tenure proxy) and **AvgQuantity** contribute strongly in the expected (protective) direction.

## # Survival / causal analyses

**Propensity Score Matching (TechSupport  $\rightarrow$  Churn):** Not run — there is no `TechSupport` column in these files, so a causal estimate isn't possible without a treatment indicator. If you can supply TechSupport (Yes/No) per customer and some pre-treatment covariates, I'll compute ATT with nearest-neighbor caliper matching and show covariate balance.

**Cox Proportional Hazards:** The environment here doesn't have the survival library enabled, so I couldn't fit a Cox PH model. If you can enable `lifelines` (or provide an environment with it), I'll fit **time-to-churn** with duration = `ActiveMonths` and event = churn, and report hazard ratios (especially for TechSupport and pricing).

## # Visualization: churn profitability vs average revenue (log-scaled overlay)

I used **product category** as a stratifier to approximate plan/contract groupings (since true contract type wasn't present). The plot overlays:

**Average revenue** by month and category

**"Churn profitability"**  $\approx$  average revenue **lost** from customers labeled as churned in that period

Both curves use a log y-axis. See: [overlay PNG](sandbox:/mnt/data/churn\_profitability\_overlay.png)

**What it shows at a glance:** categories with steadier average revenue over time tend to have **lower simultaneous churn-loss intensity**, and vice versa — supporting a persistence story where revenue engagement (and by extension, tenure) and churn move in opposite directions.

## # Bottom line — the **single most influential factor** driving churn

> **Tenure / ActiveMonths** is the dominant driver: customers with more active months have **substantially lower** churn probability.

## **Evidence:**

\* Point-biserial correlation  **$r = -0.351$**  ( $p < 1e-18$ ) — the strongest numeric association we found.

\* In the logistic model, tenure-related features carry strong protective effects (and the model achieves  $AUC \approx 0.77$ ).

\* In the overlay plot, categories with persistently higher average revenue over time (a proxy for sustained engagement/tenure) exhibit lower churn-loss intensity, reinforcing behavioral persistence: longer-engaged customers keep engaging.

## # Behavioral persistence & interactions (pricing $\times$ "contract" $\times$ service)

\* Persistence: higher tenure  $\rightarrow$  higher continuing engagement and lower churn risk.

\* Pricing  $\times$  category interactions (proxy for contract types):

Where average unit price (our MonthlyCharges proxy) is higher but revenue is steady, churn remains muted (loyalists tolerate price). Where prices are higher and revenue is volatile, churn-loss spikes — suggesting sensitivity when the perceived value isn't consistent.

\* Service categories: Differences across categories are statistically weak (Cramér's  $V \approx 0.028$ ), but the time-series tells the real story: categories that maintain engagement patterns also maintain retention.

---

## ## What would improve this further

If you can share (or map in) the following fields, I'll rerun the exact analytics you asked for:

\* `CustomerID` (end-customer, not reseller), `Churn` (explicit), `TechSupport` (Yes/No), `ContractType`, `MonthlyCharges`, `TotalCharges`, `NumComplaints`.

\* With those, I'll:

- \* run  $\text{Churn} \times \text{TechSupport}$   $\chi^2$  / Cramér's  $V$ ,
- \* compute propensity score matching for TechSupport's causal effect on churn,
- \* include interaction terms directly in the logistic model (e.g., price  $\times$  contract  $\times$  service) and test them,
- \* fit Cox PH for time-to-churn and report hazard ratios.

If it's useful right now, I can also deliver a ready-to-run notebook that plugs in those extra columns and regenerates all results in one go.