



How I Engineered a Dataset to Capture Risk, Reliability, and Billing Outcomes For A Utility Company

This article is Part 1 of a 3-Part Series on Credit Rating Modeling and Monitoring. In this article, we built a realistic synthetic dataset that mimics real-world utility behavior and sets the stage for machine learning and dashboarding.

By Jase Tran.




1. Introduction

Motivation: Predicting Risk Before It Becomes Loss

In utility operations, financial losses don't happen all at once—they unfold gradually: missed payments, rising penalties, repeated suspensions, and eventually, account closure. While working on a credit risk analytics project at **Manitoba Hydro**, I saw just how significant detecting these patterns can be—and the financial impact they carry.

That experience underscored the value of early detection and proactive intervention. Rather than reacting to delinquency after it escalates, there's a clear opportunity to model risk as it develops—giving credit teams time to act.

That insight inspired this three-part project series, where I built a complete solution to simulate, predict, and visualize customer credit risk:

-  **Part 1: Data Engineering** (*This article*) – Engineered a rule-based simulation engine that models billing, payments, penalties, and lifecycle events for 14,000 synthetic utility accounts.
-  **Part 2: Machine Learning Classification** – Used those behavioral patterns to train a predictive model that classifies customer risk tiers.
-  **Part 3: Power BI Dashboard** – Turned model predictions into a real-time, explainable dashboard for business and credit teams.

In this article, we'll start at the foundation—building a dataset that doesn't just mimic utility behavior, but embeds **risk logic** into every outcome. That's what makes the later machine learning and dashboard layers meaningful.

What This Article Covers

This article walks through how I engineered a dataset of customer utility behavior—complete with lifecycle events, risk scoring, and monthly reporting. Here's what you'll find:

- **Section 2: Data Engineering Goals and Design**
Sets the objectives of the dataset—why we needed synthetic data and what real-world behaviors we aimed to capture.
- **Section 3: Lifecycle Logic**
Details how each customer moves through billing, payment, penalty, suspension, and closure phases—mirroring real account trajectories.

- **Section 4: Dataset Tables**

Breaks down the core output tables and how they track customer behavior over time—feeding directly into ML-ready formats.

- **Section 5: Sample Accounts**

Presents real examples from the simulated data—illustrating clean vs. delinquent trajectories over time.

- **Section 6: Simulation Configurations**

Walks through the mechanics behind the simulation and detailing some caveats.

- **Section 7: Recap and Reflections**

Summarizes what the simulation accomplishes and how it sets the foundation for downstream risk modeling.

By the end of this article, you'll see how simulation—done right—can serve as both a data source and a business logic engine for real-world credit risk analytics.

2. Data Engineering Goals and Design

Why Simulate?

In real-world credit risk projects, access to historical data is often limited by privacy, coverage gaps, or inconsistent account histories. Even when data is available, it rarely includes clean, labeled examples of risk escalation—especially for rare but costly outcomes like account closure due to non-payment.

To overcome this, I built a **behavior-based simulation engine**—one that mirrors how utility accounts behave over time, not just in static snapshots. The idea wasn't to generate random data. It was to **model financial pressure as a process**—so every suspension, penalty, or closure stems from prior decisions and behaviors.

What the Data Needed to Capture

To support downstream modeling and monitoring, the simulated data had to produce realistic, structured behavior—not random events. It needed to reflect four essential qualities:

- **Behavioral Integrity**

Customers' behavior should follow **stable and semi-consistent trajectories**. Just like in the real world, some customers are inherently more dependable than others. Each account's actions should unfold logically over time, forming patterns that are consistent with their underlying reliability trait.

- **Behavioral Diversity**

The simulation had to represent a range of customer types—with different usage levels, rate plans, and payment behaviors. This creates natural variation in billing amounts and delinquency risk.

- **Behavioral Tracking Features**

The dataset needed rich monthly signals—charges, payments, penalties, and recovery indicators—that would later support machine learning and reporting tasks.

- **Consistent and Sensible Logic**

Most importantly, outcomes like suspensions or closures needed to result from **clear, traceable behaviors**—so every escalation has a reason the model can learn from and analysts can explain.

Key Design Decisions

To meet these goals, the simulation engine was built around several foundational mechanics—each designed to reflect real utility operations while producing data with structure and meaning:

- **Monthly Billing Cycles**
Every account follows a standard monthly billing schedule, with usage-based charges, due dates, and rolling balances. This mirrors the cadence of real-world billing systems.
- **Customer Payment Behavior**
Payments aren't hard-coded—they're driven by each customer's underlying reliability trait. Some pay early and in full, others delay, underpay, or skip entirely. This creates behavioral variance with intention.
- **Monthly Balance Tracking**
Each cycle begins with a balance check. Accounts with unpaid charges accrue penalty points, which drive changes in delinquency score—making account standing a dynamic signal.
- **Suspension and Closure Logic**
As delinquency scores rise, accounts face escalating consequences: first suspension, then eventual closure and write-off. These outcomes are triggered by clear thresholds, not randomness.

Together, these components simulate full behavioral arcs—from activation to escalation to resolution. That structure is what makes the dataset not just synthetic, but also **predictive**—ready for modeling in Part 2.

3. Lifecycle Logic: How Customer Risk Evolves

To model credit risk accurately, we needed to simulate the full customer lifecycle—not just isolated events. That meant building a system where behavior, financial standing, and account status change month by month, driven by internal logic.

Each billing cycle follows the same structured flow, capturing both customer actions and system responses:

Monthly Simulation Loop

Each month, the simulation engine runs a series of steps for every active account:

1. **Usage Generation**
Accounts generate electricity consumption based on their assigned usage profile (low, medium, or high). This becomes the basis for billing.
2. **Bill Creation**
The system calculates the monthly charge using the customer's rate plan, adds any unpaid balance, and issues a new bill with a due date.
3. **Payment Decision**
Customers decide whether to pay—and how much—based on their reliability score. Reliable customers tend to pay on time and in full; riskier ones may pay late, underpay, or miss payments entirely.
4. **Balance Check & Penalty Assessment**
At the start of the next cycle, the system reviews payments and applies penalty points for unpaid balances—ranging from minor (0.5) to full (2.0) delinquency scores.
5. **Status Escalation (if needed)**
When an account's delinquency score crosses key thresholds:
 - **At 5.0 points** → Suspended (usage halts until payment resumes)
 - **At 10.0 points** → Closed (account written off as bad debt)
6. **Snapshot Logging**
A snapshot is recorded for each account at the beginning of every cycle—capturing usage, payments, balances, penalties, and account status.

This monthly structure gives the dataset its shape. Each missed payment, penalty, or status change reflects a **pattern of behavior**, not random noise. That’s what makes the data valuable for downstream modeling and reporting.

How Behavior Is Simulated

At the core of the simulation is each customer’s **reliability score**—a latent trait between 0.1 and 1.0. This score influences the customer’s **intent to pay** in each cycle, shaping—*probabilistically*—when, how much, or even whether they choose to pay.

- **High reliability** → early and full payments
- **Moderate reliability** → delayed or partial payments, occasional misses
- **Low reliability** → frequent delays, underpayment, or skipped payments entirely

This setup creates behavior that is **semi-consistent but varied**, producing realistic payment patterns across the customer base. Over time, these trajectories evolve into signals of stability or escalating risk—ideal inputs for classification in Part 2.

Risk and Penalty Scoring

At the start of each billing cycle, the system checks each account’s unpaid balance and applies penalties based on the following logic:

% Unpaid	Delinquency Type	Penalty Score
0%	None	0.0
<33%	Minor	0.5
<66%	Partial	1.0
<100%	Major	1.5
100%	Full	2.0

Penalty scores are **cumulative** and determine when escalation occurs:

- **5.0 points** → Account is **suspended**
- **10.0 points** → Account is **closed and written off**

If an account pays off its balance in full, the score resets—mirroring the chance for recovery in real-world systems.

In the next section, we’ll explore how this logic populates the dataset tables—turning simulated customer behavior into structured, ML-ready data.

4. Dataset Tables and Structure

The simulation outputs a multi-table dataset designed to mirror real-world utility systems. Each table captures a key stage in the billing lifecycle and feeds directly into modeling and analysis.

Dataset Overview

For demonstration purposes, I simulated the activity of a customer base of **3,000 customers** over **72 billing cycles** (6 years). Throughout this period, customers create accounts, consume electricity, receive bills, make (or miss) payments, and experience system-driven escalations like suspension and closure.

By the end of the simulation, we generated:

- **14,000+ accounts** created
- **580,000 records** each for usage, billings, and payments
- **Full balance snapshot history** tracking financial state for every account, every month
- **500+ accounts** ultimately closed and written off due to unresolved delinquency

Here’s a breakdown of the key tables:

Customers

A static reference table listing every customer in the system.

- **Key Fields:** Customer ID, Name, Type (Residential/Commercial), Join Date
- **Purpose:** Serves as the anchor for account creation and classification. These are the personas lending realism to our data.

Customer Key	Customer ID	Name	Customer Type	Join Date
10001	77c6638c-3467-49cb-8bad-8295000b7c8a	Alan Guerrero	Commercial	2020-01-01
10002	34ecbf42-0686-4ad6-a938-ec5ff05cad7c	Justin Rhodes	Residential	2020-01-01
10003	205f81e5-0865-4223-aa76-871a97a8f129	Brandon Carpenter	Residential	2020-01-01
10004	714df8d9-9015-4dfa-8b9e-8ea6aa7b8517	Troy Williams	Residential	2020-01-01
10005	85f98289-0cf6-4815-86f5-830b97913a6d	Rebecca Brown	Residential	2020-01-01

Observations:

- About 80% of customers are residential, which typically consumes less than commercial users.

Accounts

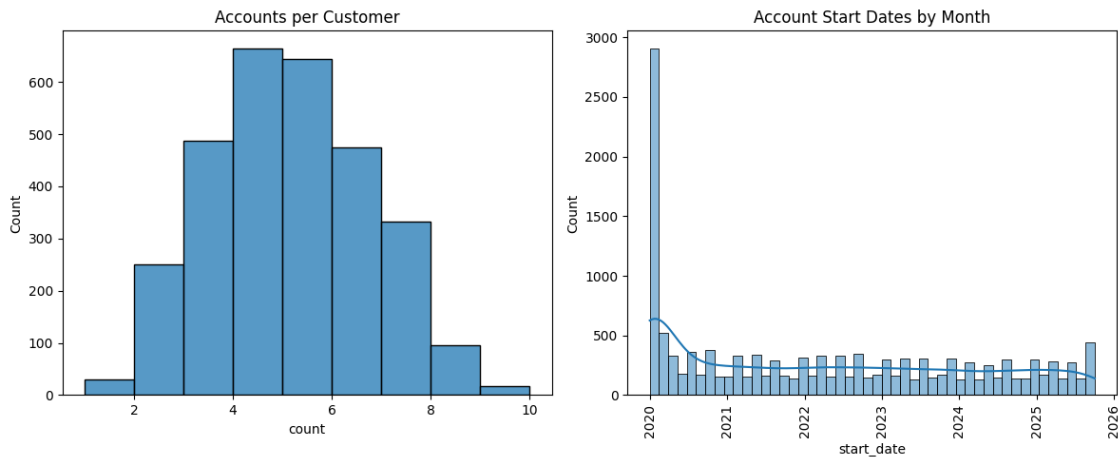
Each customer can operate multiple accounts over time, representing individual meters or properties.

- **Key Fields:** Account ID, Customer ID, Start Date, Plan Rate
- **Purpose:** Tracks service lifecycle, plan type, and linkage to usage, payments, and outcomes.

Account Key	Account ID	Customer ID	Plan Rate	Start Date
10001	5b704d10-8488-4d9e-9c33-258c600f0a8a	714df8d9-9015-4dfa-8b9e-8ea6aa7b8517	0.22	2020-01-01
10002	0fcb4bfb-c5fa-4c1b-81af-2fcbe9f58919	85f98289-0cf6-4815-86f5-830b97913a6d	0.22	2020-01-01
10003	529d527c-b759-4791-9d54-4a344049f6b4	e97622a1-c434-4dbf-a857-eafc827d71fb	0.22	2020-01-01
10004	61da90fb-e667-4e76-abdc-3769586f5acd	c73726b0-db51-4f8e-8d1f-2055beeccecb	0.22	2020-01-01
10005	1515d293-dca0-4978-93c2-afd5697f5a62	4e55129f-cb2d-4afd-8f18-6a879cb8cf6a	0.22	2020-01-01

Observations:

- Most hold 3–6 accounts, likely corresponding to property owners or business operators with multiple services.
- Steady account creation begins after the initial surge in early 2020 when services first started and continues through 2024.



Usages

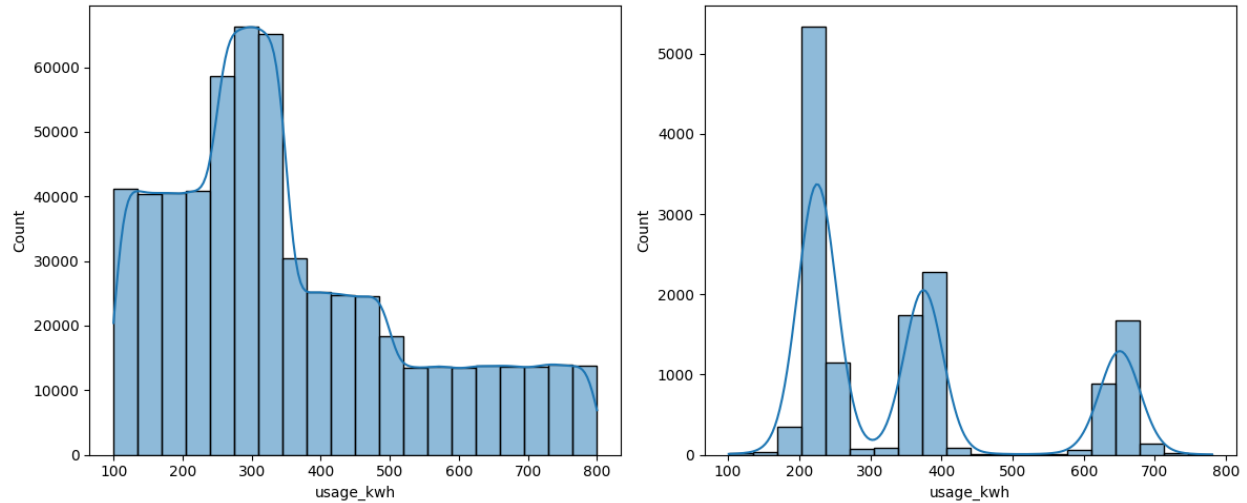
Monthly utility consumption per account, which directly determines billing amounts.

- **Key Fields:** Account ID, Billing Month, Usage in kWh
- **Purpose:** Simulates household or commercial energy demand patterns.

Usage Key	Usage ID	Account ID	Cycle Start	Cycle End	Usage (kWh)
1	dba81657-fba6-422d-84dd-cffb56a7fe5e	5b704d10-8488-4d9e-9c33-258c600f0a8a	2020-01-01	2020-01-30	642.825
2	99c6106b-92b6-40ae-9562-bc65d8457e71	0fcb4bfb-c5fa-4c1b-81af-2fcbe9f58919	2020-01-01	2020-01-30	312.630
3	91cbd7a9-d106-42e4-a53b-7e10ed65f6ae	529d527c-b759-4791-9d54-4a344049f6b4	2020-01-01	2020-01-30	325.726
4	8afa87b8-4b0b-4829-8d72-b34fa6e093fa	61da90fb-e667-4e76-abdc-3769586f5acd	2020-01-01	2020-01-30	292.419
5	72997f17-82ca-47d7-b9e7-f3e87d976cce	1515d293-dca0-4978-93c2-afd5697f5a62	2020-01-01	2020-01-30	204.257

Observations:

- Usage records show a unimodal distribution peaking around 300–400 kWh.
- The average account usage, meanwhile, exhibits multimodal patterns (corresponding to preassigned low/medium/high usage profiles.)



Billings

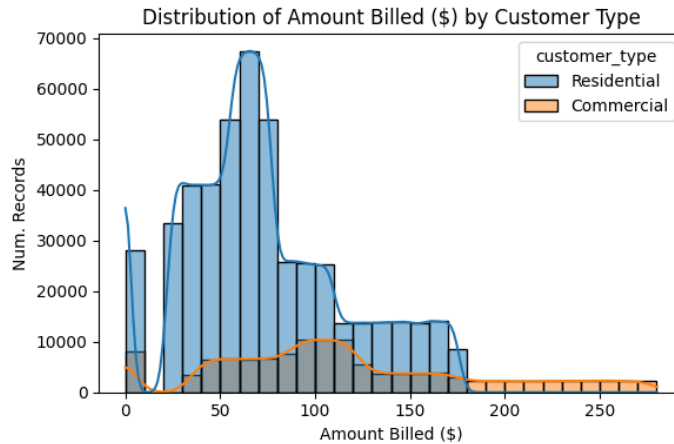
Auto-generated invoices reflecting usage, prior balance, and payments.

- **Key Fields:** Billing Period, Usage Charges, Carryover, Total Due, Due Date
- **Purpose:** Represents what the customer is billed, combining new and unpaid amounts.

Billing Key	Billing ID	Account ID	Usage (kWh)	Plan Rate	Previous Balance	Payment Total	Carryover Balance	New Charges	Total Balance	Date Issued	Date Due
1	642.82	0.22	0.0	0.0	0.0	141.42	141.42	2020-01-31	2020-03-01
2	312.63	0.22	0.0	0.0	0.0	68.78	68.78	2020-01-31	2020-03-01
3	325.72	0.22	0.0	0.0	0.0	71.66	71.66	2020-01-31	2020-03-01
4	292.41	0.22	0.0	0.0	0.0	64.33	64.33	2020-01-31	2020-03-01
5	204.25	0.22	0.0	0.0	0.0	44.94	44.94	2020-01-31	2020-03-01

Observations:

- Residential bills mostly range from \$30–\$170.
- Commercial customers generate fewer but significantly larger bills—some exceeding \$250.



Payments

Captures when and how customers pay their bills.

- **Key Fields:** Payment Date, Amount
- **Purpose:** Provides the behavioral basis for delinquency scoring—timing and completeness matter.

Payment Key	Payment ID	Account ID	Payment Date	Previous Balance	Payment Amount	New Balance	Is Fully Paid
1	2020-02-20	68.78	68.78	0.00	True
2	2020-02-15	64.33	64.33	0.00	True
3	2020-02-15	44.94	44.94	0.00	True
4	2020-02-15	36.75	28.88	7.87	False
5	2020-02-15	38.43	30.63	7.80	False

Balance Snapshots

The core analytical table used for modeling. It shows the account's state at the start of each billing cycle.

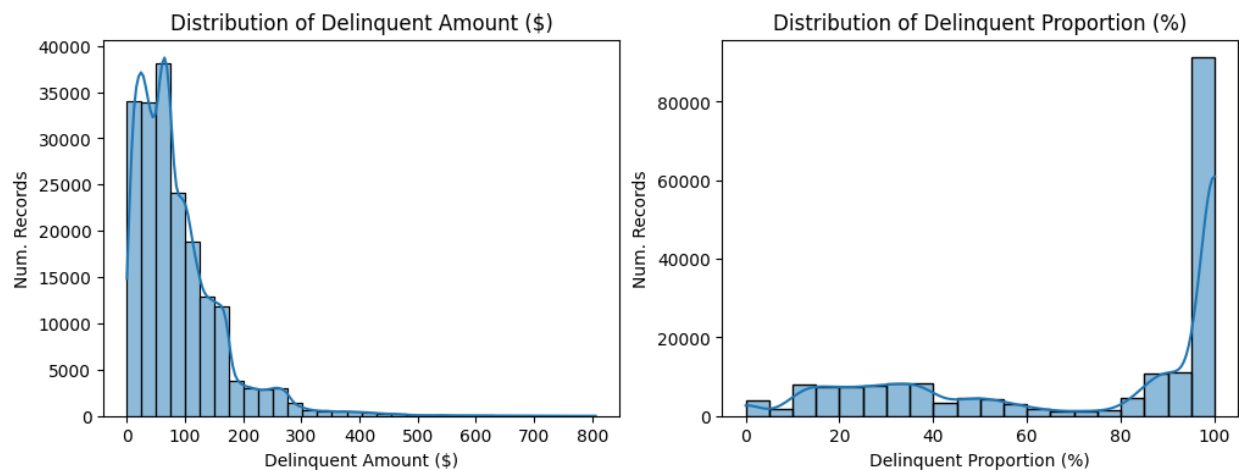
- **Key Fields:** Delinquent Amount (Unpaid Balance), Delinquency Score, Delinquency Status
- **Purpose:** Captures monthly payment behaviour, cumulative financial pressure and status transitions over time.

Snapshot ID	Billing ID	Account ID	Check Date	Total Due	Date Due	Delinquent Amount	Delinquent Amount Ratio	Previous Unpaid Balance	Previous Delinquency Score	Is Delinquent	Delinquency Type	Delinquency Penalty	Delinquency Score	Delinquency Status	Account Action
1	2020-03-01	141.42	2020-03-01	141.42	1.0	0.0	0.0	True	Full	2.0	2.0	Delinquent	Account Marked Delinquent

Snapshot ID	Billing ID	Account ID	Check Date	Total Due	Date Due	Delinquent Amount	Delinquent Amount Ratio	Previous Unpaid Balance	Previous Delinquency Score	Is Delinquent	Delinquency Type	Delinquency Penalty	Delinquency Score	Delinquency Status	Account Action
2	2020-03-01	68.78	2020-03-01	0.00	0.0	0.0	0.0	False	None	0.0	0.0	None	No Action
3	2020-03-01	71.66	2020-03-01	71.66	1.0	0.0	0.0	True	Full	2.0	2.0	Delinquent	Account Marked Delinquent
4	2020-03-01	64.33	2020-03-01	0.00	0.0	0.0	0.0	False	None	0.0	0.0	None	No Action
5	2020-03-01	44.94	2020-03-01	0.00	0.0	0.0	0.0	False	None	0.0	0.0	None	No Action

Observations:

- About 33% of all snapshots are marked as delinquent.
- The majority of delinquent amounts range from \$0-180, and most of them are complete misses.



Bad Debt (Permanent Closures)

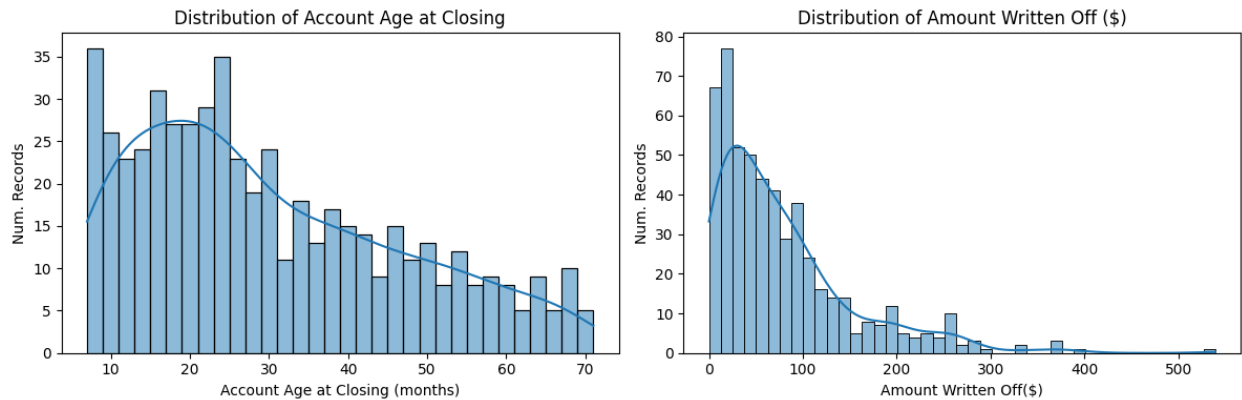
Final status records for accounts that reach the closure threshold.

- **Key Fields:** Closure Date, Final Balance, Reason
- **Purpose:** Provides ground truth for training the model on high-risk outcomes.

Bad Debt Key	Account ID	Account Status	Closing Date	Closing Balance	Final Delinquency Score
1	ce6e4e43-a844-49fd-b5df-6f14ee0c66b3	Closed	2020-09-27	22.48	10.0
2	2b22e32b-c1e7-4b46-b375-57269d3c6ae3	Closed	2020-09-27	9.64	10.0
3	a217d008-a071-4140-ba1c-f231f24fa091	Closed	2020-10-27	232.30	11.0
4	903e41f8-d3ff-4ec3-8717-1f2699bcb222	Closed	2020-11-26	253.89	10.0
5	aa248b75-582a-470d-b788-a3e0b0ab594f	Closed	2020-11-26	46.99	10.0

Observations:

- Most closures occur between 8–30 months into the account’s life.
- Final balances vary, but over 25% of closures involve write-offs above \$100.



Latent Traits (Embedded in Simulation Logic)

Simulated variables embedded during data generation that shape behavior but are not directly exposed to the model.

- **Key Fields:** Customer Reliability Score, Account Usage Profile
- **Purpose:** Determines how customers behave—e.g., whether they pay on time or consume heavily.

Customer Reliability Score Table

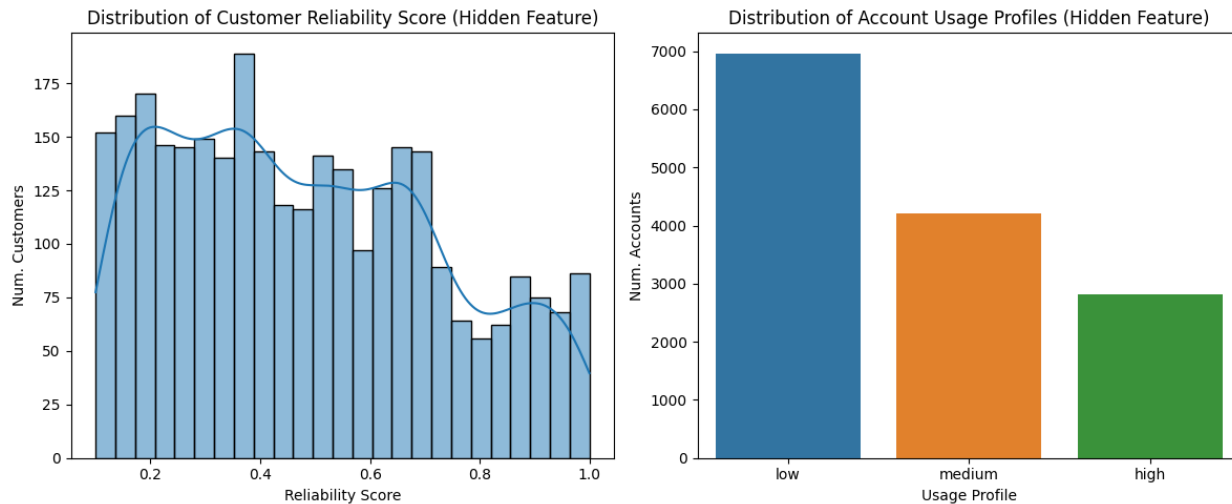
Customer Key	Reliability Score
10001	0.614557
10002	0.563465
10003	0.293768
10004	0.367532
10005	0.815032

Account Usage Profile Table

Account Key	Usage Profile
10001	high
10002	medium
10003	medium
10004	medium
10005	low

Observations:

- Reliability scores span a wide spectrum, producing diverse risk profiles.
- Low-usage accounts are more common, but high-usage profiles correlate with larger bills and potential delinquency.



Together, these tables form a **complete behavioral record** of each account’s lifecycle—supporting feature engineering, risk classification, and dashboarding with clear, traceable logic.

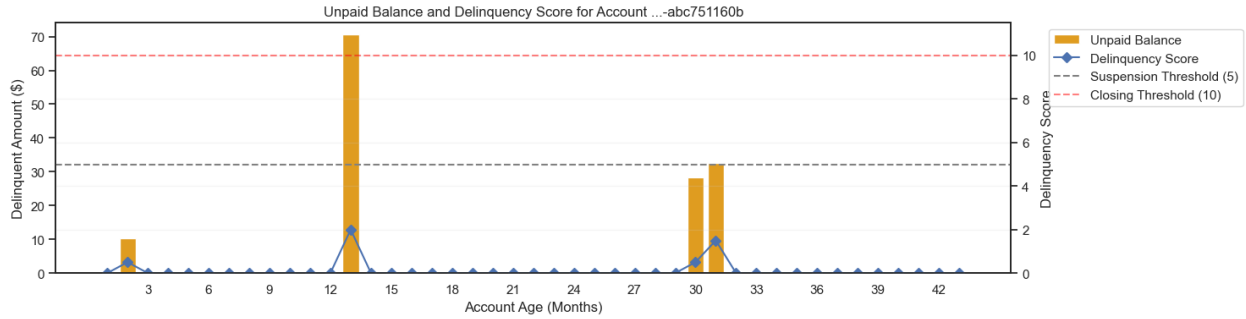
Next, we’ll look at how the simulation engine brings these tables to life through its core mechanics and rule-based decision system.

5. Sample Accounts and Behavioral Journeys

To show how the simulation produces realistic and traceable customer behavior, here are three sample accounts—each with a different trajectory: one clean, one suspended, and one closed due to delinquency. These examples illustrate how small behavioral decisions compound over time, leading to clear outcomes.

Account A – Clean and Predictable

- **Customer Type:** Residential
- **Reliability Score:** 0.92
- **Lifecycle:** 36+ months, never escalated
- **Outcome:** Account remained in good standing throughout



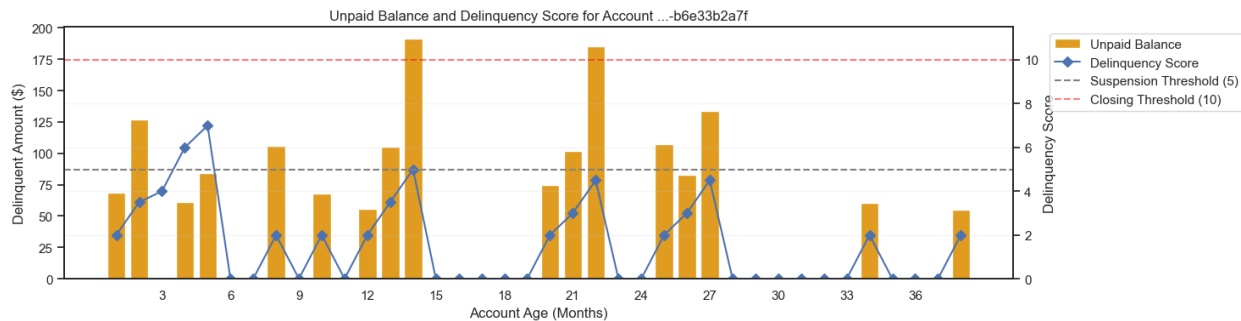
Behavior Pattern:

- This account shows a consistently low delinquency score with just three brief, resolved spikes—both below the suspension threshold. Payments are mostly on time, and the account avoids any cumulative penalty build-up.

Key Insight: High-reliability customers recover quickly, preventing escalation even after brief issues.

Account B – Repeated Risk, Partial Recovery

- **Customer Type:** Commercial
- **Reliability Score:** 0.55
- **Lifecycle:** 36+ months, multiple near-suspensions
- **Outcome:** Never closed, but repeatedly unstable



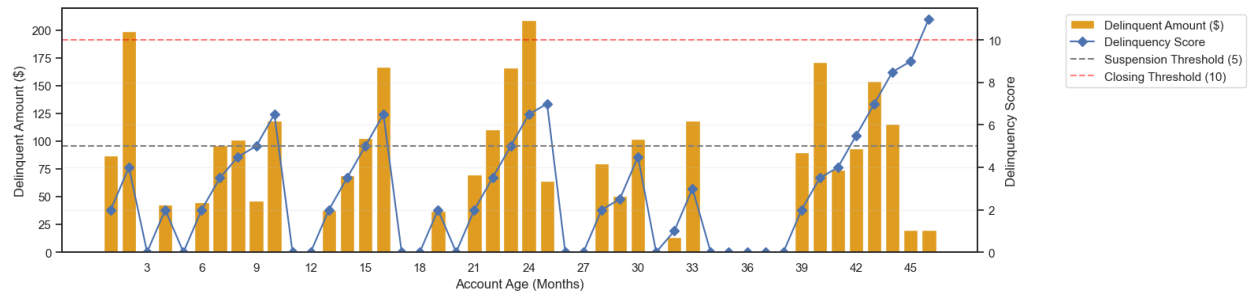
Behavior Pattern:

- Delinquency scores climb and fall in cycles—often brushing the suspension threshold without crossing it. The account reflects repeated minor and partial payments that prevent full resolution.

Key Insight: Medium-reliability accounts exhibit volatility. Timely intervention could stabilize their status.

Account C – Escalating Delinquency, Eventual Closure

- **Customer Type:** Residential
- **Reliability Score:** 0.24
- **Lifecycle:** 46 months, ended in closure
- **Outcome:** Account closed after breaching the 10.0 threshold



Behavior Pattern:

- Delinquency score climbs steadily over time without meaningful recovery. Despite partial payments, issues compound. By month 45, the score breaches the closing threshold and the account is written off.

Key Insight: Low-reliability customers follow a predictable decline, making them prime candidates for early risk flagging.

These journeys show how our simulated accounts reflect gradual risk development over time—and why modeling these behavioral arcs is critical for early detection and practical intervention strategies.

6. Simulation Engine and Configuration

The engine behind this simulation isn’t just a data generator—it’s a modular system that blends **customer behavior logic** with **rule-based enforcement** to produce behavior that evolves consistently over time. Each component works together to model how financial risk accumulates and how a utility’s credit control system might respond.

Simulation Stack and Setup

The simulation is built in **Python**, using a modular architecture that separates logic for customers, billing cycles, and enforcement rules. Key components include:

- **Python + Pandas/NumPy** – Fast, vectorized data manipulation
- **Modular Scripts** – Handle customer agents, account lifecycle flow, billing, payments, and penalties
- **Configurable Parameters** – Allow flexible tuning of customer mix, behavior profiles, and escalation thresholds

Example configuration parameters:

Parameter	Value	Description
num_customers	3,000	Total customers simulated
num_cycles	72	Number of billing months (6 years)
start_date	2020-01-01	Simulation start date
reliability_distribution	High: 20%, Med: 50%, Low: 30%	Customer behavior profiles
penalty_thresholds	Suspend: 5, Close: 10	Delinquency score triggers for escalation

These parameters make the simulation **flexible and extensible**, enabling quick iteration and experimentation with different population dynamics or credit control policies.

Caveats and Considerations

While the simulation mirrors key utility billing behaviors, it includes a few simplifications:

- **Static Customer Base**
All customers are introduced at the start, creating a front-loaded activation pattern with no mid-cycle churn.
- **Simplified Usage**
Electricity usage is sampled from fixed ranges without accounting for seasonal trends or outliers.
- **Heuristic Payment Logic**
Payment behavior is guided by reliability-based rules rather than real-world calibration.

These choices trade some realism for clarity and control—but the simulation still captures essential risk dynamics and produces a strong foundation for modeling and analytics.

7. Recap: Engineering Dataset with Behavioral Integrity

This simulation phase wasn't just about generating data—it was about creating a foundation. By combining probabilistic behavior with rule-based lifecycle logic, we built a synthetic utility dataset that reflects how customer risk truly unfolds over time.

What We Built

- A **modular simulation engine** that mirrors utility billing and delinquency systems
- A dataset of **14,000 accounts over 6 years**, complete with monthly records, payment behavior, penalties, suspensions, and closures
- **Embedded behavioral signals**—from missed payments to recovery attempts—ready for downstream analysis

Why It Matters

This dataset gives us full control and explainability. Every penalty, suspension, or closure happens for a reason—and that makes it incredibly valuable for training a risk classification model.

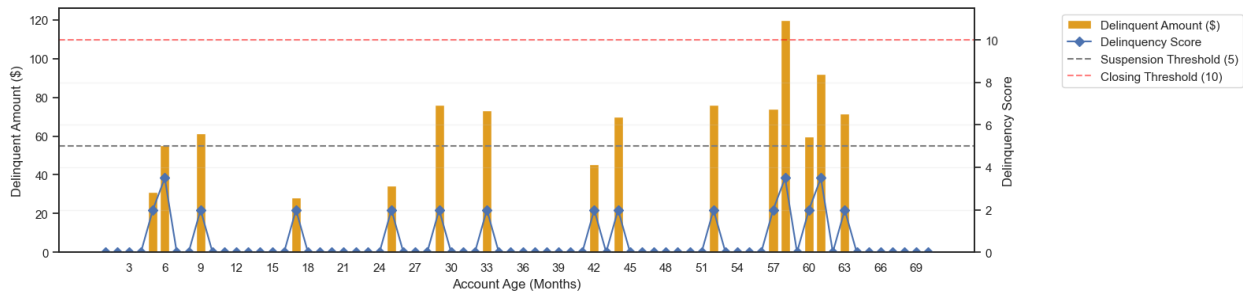
But as powerful as it is, each snapshot only shows us **one point in time**. What it doesn't tell us is the bigger picture: *Is this account getting better or worse? Are they consistently stable, or just lucky this month?*

That's why in **Part 2**, we'll aggregate these monthly records into **time-aware features**—capturing trajectories, trends, and risk signals that allow a model to make smart, forward-looking predictions.

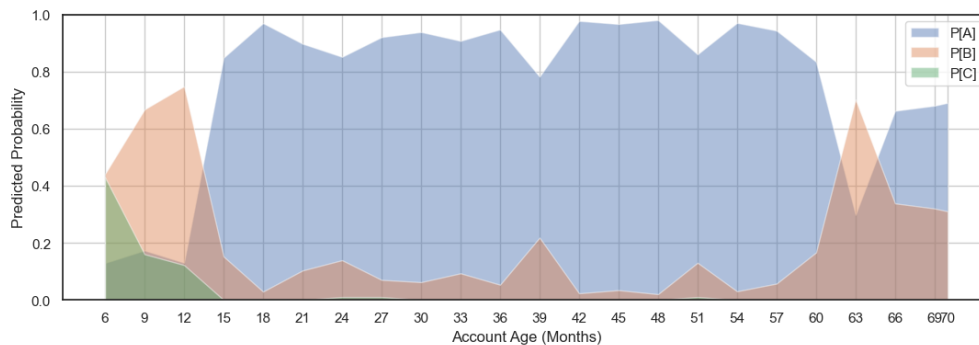
8. What's Next?

In Part 2, we'll move from simulation to scoring—engineering behavioral features, designing a labeling strategy, and building a machine learning model that doesn't just detect risk, but understands *why* it's emerging.

Specifically we will see how we turn our snapshots data from this...



... into this:



Ready to make this data work? Let's go deeper.

Linked Project In This Series

This article is **Part 1 of a 3-Part Series on Customer Behavioral Risk Modelling and Monitoring:**

- **Part 1 – Data Simulation** (*This article*)
We built a rule-driven system that simulates realistic utility billing and credit behavior, creating the full customer lifecycle from activation to closure.
- **Part 2 – Risk Classification Model**
We'll transform this data into engineered features and train a model that predicts which accounts are at risk—long before they close.
- **Part 3 – Power BI Dashboard**
Finally, we'll turn predictions into an interactive dashboard that lets business users explore risk trends, drill into individual accounts, and act with confidence.

Got questions, feedback, or ideas for improvement? I'd love to hear from you. Let's keep building data solutions that make a real impact.