



# How I Created A Robust Credit Rating System For A Utility Company (Using Machine Learning)

*This article is Part 2 of a 3-Part Series on Credit Rating Modelling and Monitoring. In Part 1, we created a simulated utility billing system and generated a dataset that resembles real-world customer billing and payment behavior. In this article, we use that behavioral history to train a predictive model that classifies risk before accounts become closed due to delinquency.*

By Jase Tran.

---

## 1. Introduction

### Motivation: Predicting Risk Before It Becomes Loss

In utility operations, financial losses rarely happen overnight. More often, they follow a pattern—subtle at first, then progressively clearer: missed payments, rising penalties, suspended services, and finally, write-offs. While working as a Data Analyst at **Manitoba Hydro**, I saw that the key to mitigating these losses wasn't just detecting delinquency—it was predicting it early enough to act.

That insight inspired this three-part project series. Together, the articles walk through how I built a full-cycle solution to simulate, score, and surface customer credit risk:

- ✅ **Part 1: Data Engineering** – Created a simulated utility billing system and dataset that mimics real-world customer billing and payment behavior
- 🔄 **Part 2: Machine Learning Classification – (This article)** Built a model that classifies account risk based on behavioral patterns
- 🔄 **Part 3: Power BI Dashboard** – Visualized risk in a practical, explainable report for operations teams

Each phase builds on the last. In Part 1, we modeled the lifecycle of 14,000 synthetic accounts—including usage, payments, penalties, suspensions, and closures—using a rule-based engine that captures how risk emerges over time.

This article, Part 2, shows how we used that behavioral history to train a machine learning model that flags accounts likely to default before they actually do.

### What This Article Covers

We'll walk through the full journey of building a behavior-first risk classification model—starting with simulated billing data and ending with explainable predictions. Here's what you'll find:

- **Section 2: Dataset Foundation**  
Recaps the simulated dataset from Part 1, focusing on how rule-based billing behavior sets the stage for machine learning.
- **Section 3: Feature Engineering**  
Shows how we transformed raw billing snapshots into behavioral features—capturing trends like delinquency frequency, penalty accumulation, and recovery patterns.
- **Section 4: Labeling Strategy**  
Defines our three-tier (A–B–C) risk classes using both final outcomes and behavioral clustering—ensuring our model learns from meaningful risk signals.
- **Section 5: Modeling and Training**  
Walks through how we selected snapshots for training, built a Random Forest model, and optimized for both accuracy and real-world usability.
- **Section 6: Model Evaluation**  
Analyzes performance using confusion matrices, prediction confidence, and behavior-over-time trends.
- **Section 7: Explainability**  
Breaks down how we used SHAP values and feature importance to explain predictions—bridging the gap between data science and business teams.
- **Section 8: Case Studies**  
Walks through real examples of each risk class, showing how the model interprets behavior and flags concern early.
- **Section 9: Recap and Reflections**  
Summarizes the pipeline and insights from this phase, and how it leads into the final dashboard solution in Part 3.

By the end, we'll have a working risk model that doesn't just sort accounts into categories—but tells a story about why each one matters, and sets the stage for the dashboard in Part 3.

---

## 2. Dataset Foundation: Simulated Behavior with Real-World Logic

In Part 1, we engineered a synthetic dataset that simulates six years of monthly billing behavior for 14,000 utility accounts. Each account follows a behavioral lifecycle driven by rule-based logic—meaning every missed payment, penalty, or account closure happens for a reason.

The billing system works by reviewing each account's balance at the start of every cycle and applying penalty points for unpaid amounts. Minor payment misses add 0.5 points, while missed or severely delayed payments add up to 2.0. As delinquency points accumulate, accounts escalate automatically:

- **5.0 points** → Suspended
- **10.0 points** → Closed

At the core of the dataset is the **Balance Snapshots** table, which captures the state of each account at the start of every billing cycle, including:

- **Delinquent Amount and Ratio** – Tracks how much of the current bill remains unpaid, expressed both as a raw amount and as a percentage of the total due.
- **Delinquency Type and Penalty** – Categorizes delinquency based on severity (Minor, Partial, Major, Full), determined by the unpaid ratio. Each type maps to a penalty score from **0.5 to 2.0**, which contributes to overall risk.
- **Delinquency Score** – A cumulative index of financial strain, calculated as the **sum of all penalty points** since the last clean billing cycle. It reflects how prolonged or severe an account's issues have been.
- **Delinquency Status** – Indicates the operational state of the account: **None, Delinquent, Suspended, or Closed**, depending on delinquency score thresholds and payment recovery.

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
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

## Dataset Overview

For this project series, we simulated a starting population of **3,000 customers** over **72 billing cycles** (6 years). Over time, these customers opened multiple accounts, consumed energy, received bills, made payments, and experienced consequences for delayed or missed payments.

The simulation generated:

- **14,000+ accounts**
- **580,000+ records each** for usage, billing, and payments
- **A complete history** of balance snapshots across the lifecycle of each account
- **500+ final closures**, representing accounts written off due to unresolved delinquency

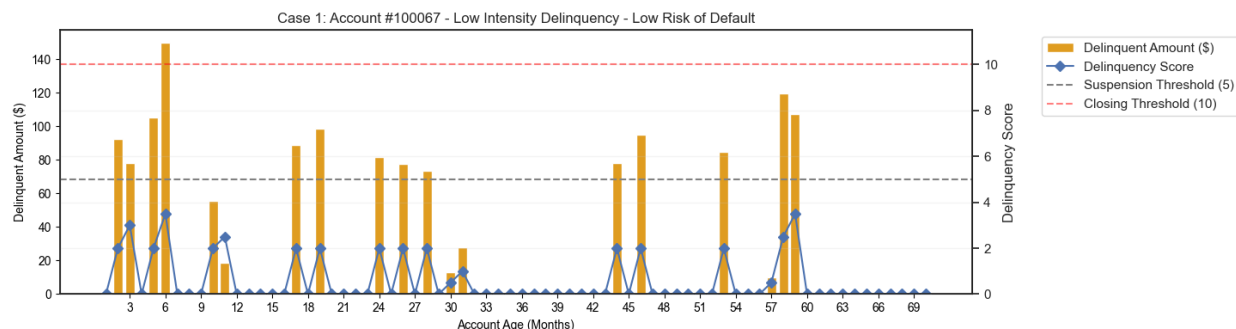
With this volume and richness, the dataset provides a robust base for both modeling and evaluation.

## Sample Accounts: Patterns That Signal Risk

The dataset contains a wide range of customer behaviors—some stable, others risky. Below are two illustrative examples:

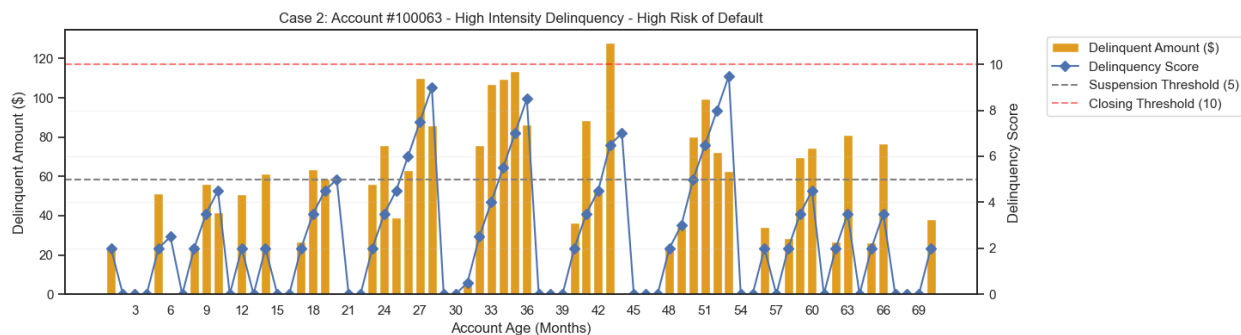
- **Case 1: Low-Intensity Delinquency – Low Risk of Default:**

This account shows occasional missed payments early on, but consistently recovers. Delinquency scores remain low and never breach escalation thresholds. The overall trajectory is stable.



- **Case 2: High-Intensity Delinquency – High Risk of Default:**

This account exhibits frequent, prolonged delinquency with escalating penalty scores. The customer repeatedly misses payments, pushing the account toward suspension and eventual closure.



To build a predictive model, we needed to turn these monthly records into something more meaningful—features that describe trends, consistency, and deterioration over time. In the next section, we’ll walk through how we engineered those behavioral signals.

---

## 3. Feature Engineering: Quantifying Behavioral Risk

### From Snapshots to Storylines

A single snapshot tells you what an account looks like at a point in time. But delinquency isn’t static—it evolves. So instead of relying on isolated records, we engineered features that summarize **how each customer got there**.

For every modeling snapshot, we looked back at the account’s history and calculated trends like:

- **Delinquency Rate** – % of billing cycles with missed or partial payments
- **Suspension Frequency** – How often the account was temporarily frozen
- **Max Penalty Score** – Peak accumulated score from repeated delinquencies
- **Recent Patterns** – How the account is performing in the recent windows (3,6,12 months)

We normalized many of these features to account for account age, ensuring fairness across newer and older accounts. For example:

$$\begin{aligned}\text{Delinquency Rate} &= \# \text{ Delinquent Months} \div \text{Account Age (Months)} \\ \text{Suspension Rate} &= \# \text{ Suspended Months} \div \text{Account Age (Months)}\end{aligned}$$

This let the model compare consistency and risk trends across the full portfolio, regardless of how long an account had been active.

### Why We Ignored Dollar Amounts

We intentionally **excluded monetary values** from the feature set. While they reflect financial magnitude, they can overshadow behavioral nuance. Our goal was to model **how** customers behave—not how much they owe—so the model would learn patterns that generalize across account sizes and billing tiers.

#### *Final Engineered Dataset*

Snapshot Key	Account Age	Max Delinquency Score	Avg. Penalty Per Incident	Delinquency Rate	Suspension Rate	Active Delinquency Rate	Avg. Penalty	Delinquency Streak Length 1M Rate	Delinquency Streak Length 2-3M Rate	Delinquency Streak Length 4M+ Rate	Payment Full Miss Rate	Payment Major Miss Rate	Payment Partial Miss Rate	Payment Minor Miss Rate
216621	16	4.5	1.722	0.562	0.000	0.562	0.968	0.250	0.125	0.000	0.375	0.125	0.000	0.062
281483	26	3.5	1.650	0.384	0.000	0.384	0.634	0.230	0.076	0.000	0.269	0.038	0.000	0.076
50048	1	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
313125	52	7.5	1.592	0.519	0.038	0.480	0.826	0.096	0.134	0.019	0.269	0.115	0.096	0.038
56279	9	3.0	1.750	0.444	0.000	0.444	0.777	0.222	0.111	0.000	0.333	0.000	0.111	0.000

---

## 4. Labeling Strategy: Defining What “Risky” Means

Before a model can learn, we have to define what it should learn *to recognize*. In our case, the challenge wasn't just to detect accounts that had already failed—but to flag those showing signs of trouble *before* they reach that point.

This means our label strategy needed to capture not just outcomes, but the behavioral patterns leading up to them.

## Starting Point: Account Closure as a Red Flag




We began with a concrete indicator: **account closure due to unpaid debt**. It's an unambiguous failure point and gives us a natural definition of high risk.

But that signal alone isn't enough:

- Some accounts remain active despite serious, repeated delinquencies.
- Others close abruptly with minimal warning signs.
- A simple binary label (closed vs. active) would miss important shades of risk.

## A–B–C Risk Classes: Reflecting Real-World Complexity

To model risk with more nuance, we designed a **three-tier classification system**:

-  **Class C – High Risk**: Accounts that were closed due to delinquency, or showed sustained behaviors that nearly reached closure thresholds.
-  **Class B – Medium Risk**: Accounts with mixed behavior—some issues, but not consistently severe or trending toward closure.
-  **Class A – Low Risk**: Accounts with strong, consistent payment behavior—rarely delinquent, rarely penalized, never suspended.

These classes give our model a more practical, business-aligned understanding of risk—useful not just for prediction, but for prioritization in operational workflows.

## Labeling Method: Hybrid of Rules and Clustering

We created these labels using a **hybrid (rule + clustering) policy**:

- **Class C**: All accounts that ended in closure, plus a set of active accounts showing near-closure behavior (e.g., prolonged high penalties or repeated suspensions).
- **Class A/B**: Among remaining active accounts with at least 24 months of history, we applied clustering on key delinquency metrics to separate clean vs. inconsistent behavior.

This allowed us to capture behavior-based risk while preserving label quality—even in a synthetic environment.

To support lifecycle-aware modeling, we propagated each account's **final label** to all earlier snapshots—teaching the model not just how accounts *end*, but how that outcome *develops* over time.

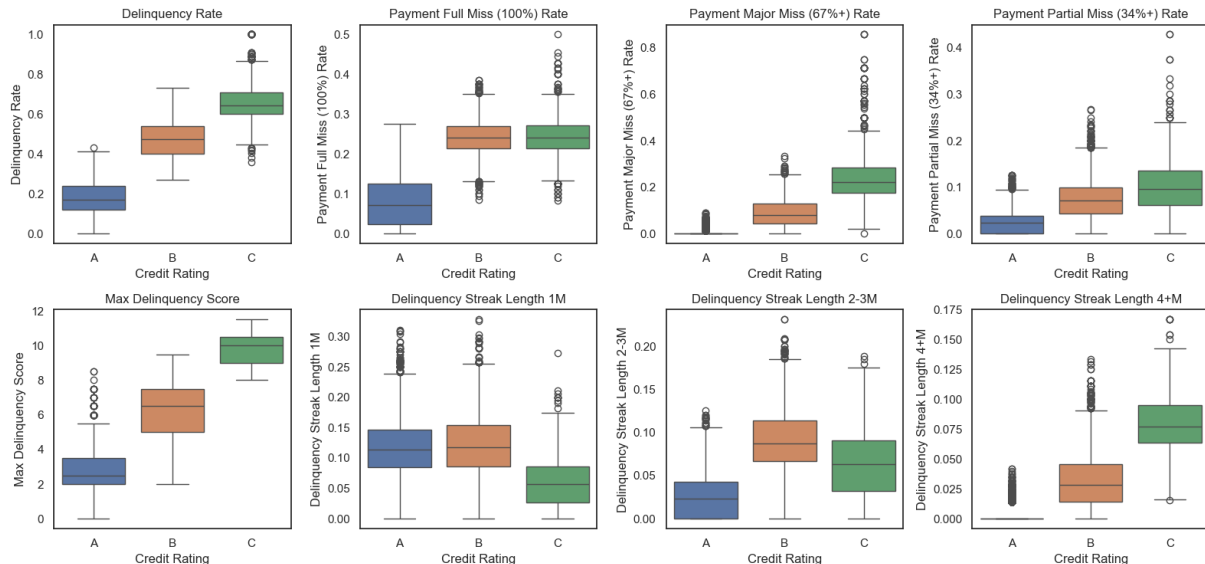


## How Our Label Policy Separates Classes

Our labeled credit classes reflect clear behavioral differences:

- **Delinquency Rate:** Class A accounts show low, infrequent delinquency. Class C accounts are consistently overdue. Class B falls between—irregular but notable issues.
- **Missed Payment Severity:** Full and major misses increase across  $A \rightarrow B \rightarrow C$ , with Class C showing the highest failure rates.
- **Max Delinquency Score:** Class A rarely accumulates penalties. Class B sees moderate escalation. Class C often exceeds suspension or closure thresholds.
- **Delinquency Streaks:** Class C accounts have the longest sustained delinquency. Class A typically resolves quickly; Class B shows mixed patterns.

Distributions of Key Selected Features per Credit Rating Class



These differences validate our labeling policy—capturing both clear risk extremes and borderline cases where early action is key.


---

## 5. Modeling and Training: Learning from Behavior

With our features engineered and labels assigned, we turned to model development. The goal wasn't just to achieve high accuracy—but to build a classifier that mimics how analysts think: looking at patterns over time and gauging whether behavior is improving, stable, or deteriorating.

### Selecting Snapshots for Training

Even though our dataset contains thousands of monthly snapshots, not all are equally useful for training. To create a learning environment that mirrors how the model would be used in practice, we applied careful sampling criteria:

- **Closed Accounts:** We included all snapshots—these show clear high-risk behavior across the lifecycle.
- **Active Accounts:** Only those with at least 24 months of history, to ensure sufficient behavioral context.
- **Timing Filter:** Snapshots must be taken at least 6 months into the account lifecycle, to avoid noise from early instability.
-  **Temporal Spacing:** We sampled one snapshot every 3 months (e.g., Month 6, 9, 12...), to reduce redundancy.
- **Excluding Final Snapshots:** Final records often include outcome clues. We removed them to avoid leakage.
- **Balanced Sampling:** To avoid class imbalance bias, we applied random under-sampling—ensuring equal training exposure across Classes A, B, and C.

This approach yielded a well-rounded training set that reflects how the model will operate in practice: identifying risk in accounts with partial, unfolding histories.

### Model Selection: Random Forest

We chose a **Random Forest Classifier** for its balance of performance and interpretability. Benefits included:

- Handles non-linear feature interactions
- Tolerant of outliers and mixed data types
- Provides feature importance metrics for explainability
- Requires minimal tuning—letting us focus on data quality

Given our structured features and clearly defined labels, Random Forest offered a strong baseline that supported both performance and business transparency.

---

## 6. Model Evaluation and Interpretation



Once trained, we evaluated the model not just on statistical metrics—but on its ability to support real decisions. Evaluation focused on three core questions:

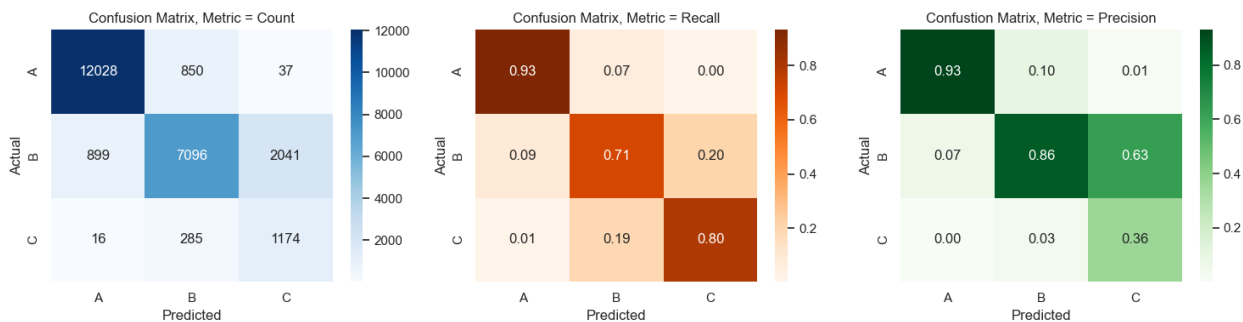
- **Class-wise Performance:** How well does the model distinguish between low, medium, and high-risk accounts?
- **Snapshot Age Impact:** How early can the model make reliable predictions?
- **Confidence Calibration:** Do higher-confidence predictions lead to better accuracy?

### Class-wise Performance: Confusion Matrix

The confusion matrices below show the model’s class-wise performance in terms of count, recall, and precision.

- **Low Risk (A):** High recall (93%) and precision (93%)—very few false positives or negatives.
- **Medium Risk (B):** Moderate recall (71%) and strong precision (86%)—captures many but not all borderline cases.
- **High Risk (C):** Good recall (80%) but lower precision (36%)—model flags risky accounts well but sometimes confuses with borderline B cases.

This reflects the inherent ambiguity in Class B and partial overlap with Class C.

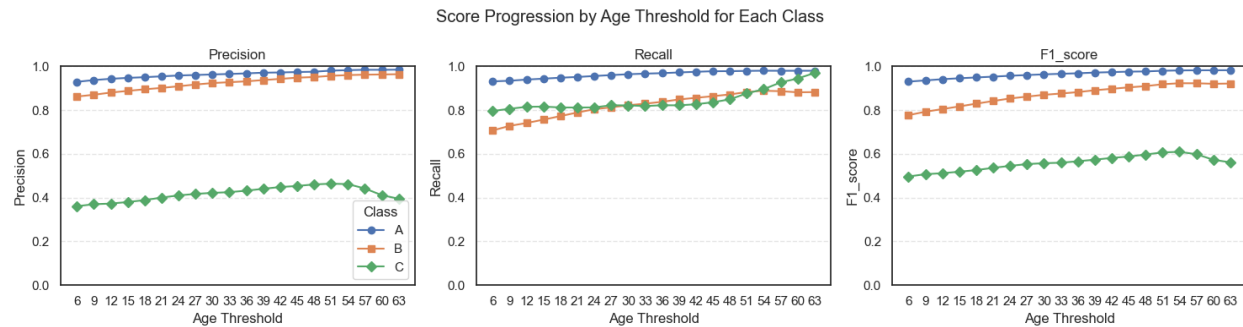


### Snapshot Age: Performance Improves Over Time

As account age increases, performance improves across all metrics:

- **Precision:** Increases steadily, especially for Classes A and B
- **Recall:** Strong early on for Class A; improves for B and C as more history accumulates
- **F1 Score:** Rises consistently with age—showing that the model benefits from behavioral depth

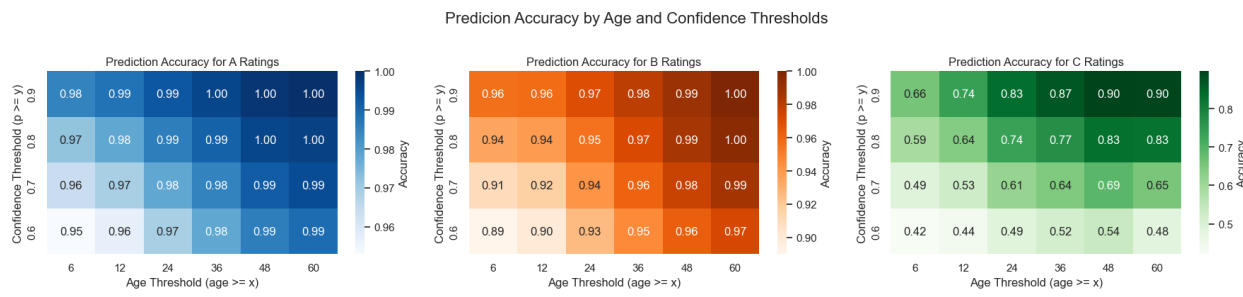
By 36-48 months, the model achieves high balance across precision, recall, and F1 for all classes.



## Prediction Confidence: Accuracy Scales with Certainty

For each prediction, the model outputs class probabilities—allowing us to assess both label and confidence. When grouped by minimum confidence levels:

- Predictions with  $\geq 0.8$  confidence are highly accurate:
  - ~94–100% for A and B ratings
  - ~75–90% for C ratings at higher age thresholds
- Accuracy scales with both confidence and account age—validating the model’s reliability under stricter thresholds



# 7. Explainability: Opening the Black Box

For our risk model to be usable in operations, transparency is essential. It’s not enough to predict risk—we need to show *why* an account was flagged. To support this, we added two layers of explainability: global and local.

## Global Feature Importance: What the Model Prioritizes

We used SHAP values to measure the average impact of each feature across all predictions.

**Top Features** included:

- Total Penalty Pts
- Delinquency Rate

- **Payment Major Miss Rate**
- **Suspension Rate**
- **Payment Full Miss Rate**

These features clearly align with business intuition—accounts with frequent penalties, suspensions, and missed payments are more likely to be high-risk. The model's priorities confirm that it learned meaningful, interpretable signals.

## Local Explanations: Why Each Account Was Rated That Way

For individual predictions, we used SHAP to break down the contribution of each feature. To illustrate how the model explains individual predictions, let's look at **Snapshot ID: 305692**, which was classified as **Class C (High Risk)**.

Three key features drove this outcome:

- **count\_suspension\_norm = 0.17 → SHAP +0.09**  
Frequent service suspensions strongly signaled instability and contributed heavily to the high-risk rating.
- **total\_penalties\_norm = 0.88 → SHAP +0.08**  
The account accumulated significant penalties, reinforcing the pattern of non-compliance.
- **total\_delinquencies\_norm = 0.58 → SHAP +0.07**  
Repeated delinquency incidents further elevated the likelihood of default.

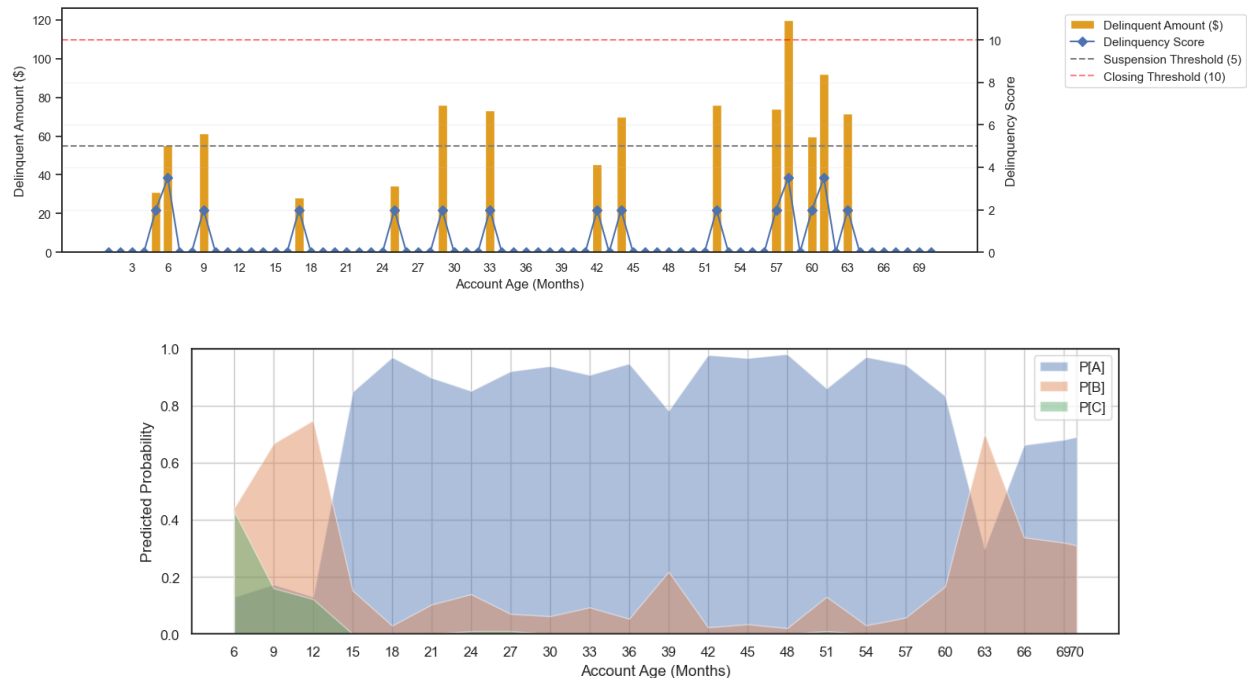
Each of these behaviors *promoted* the model's confidence in assigning a high-risk label. By breaking down the prediction into feature-level impacts, analysts gain clear visibility into *why* the account was flagged—enabling informed, traceable decision-making.

## 8. Case Studies: Risk Classes in Action

To see how the model behaves in practice, let's walk through an example from each risk class. These case studies illustrate how behavioral signals, risk probabilities, and model confidence evolve over time.

### **Account A – Low Risk**

- **Status:** Active
- **Account Duration:** 72 months
- **Delinquency History:** Intermittent, low-severity delinquency events with full resolution; no suspensions or closures
- **Prediction Pattern:** Consistently predicted as **Class A**, with probability  $\geq 85\%$  for most of the account's life



### What the Model Saw:

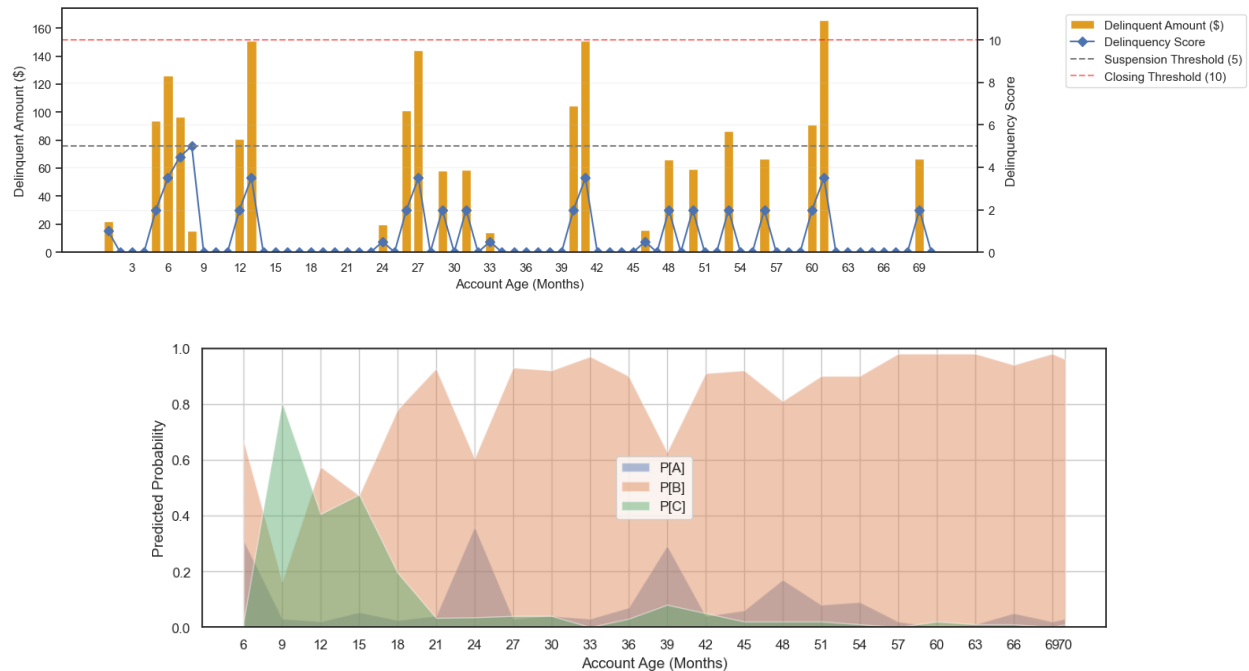
- Short-lived delinquencies quickly recovered
- Delinquency score stayed well below suspension and closure thresholds
- No escalations, even during occasional spikes in owed amount

### Interpretation:

- This is a resilient, stable account. While there were some isolated issues, they were low in intensity and never escalated. The model correctly identified this as a reliable customer—not in need of intervention.

## Account B – Medium Risk

- **Status:** Active
- **Account Duration:** 72 months
- **Delinquency History:**  
Frequent delinquency events throughout the account's life—often moderate to high in dollar amount. While the account avoided closure, it repeatedly approached the suspension threshold, signaling recurring financial strain.
- **Prediction Pattern:**  
The model steadily predicted Class B for most of the account's lifecycle, with probability rising above 85% from around month 24 onward. Class C briefly emerged early, but never dominated, and Class A faded quickly as risk signs accumulated.



### What the Model Saw:

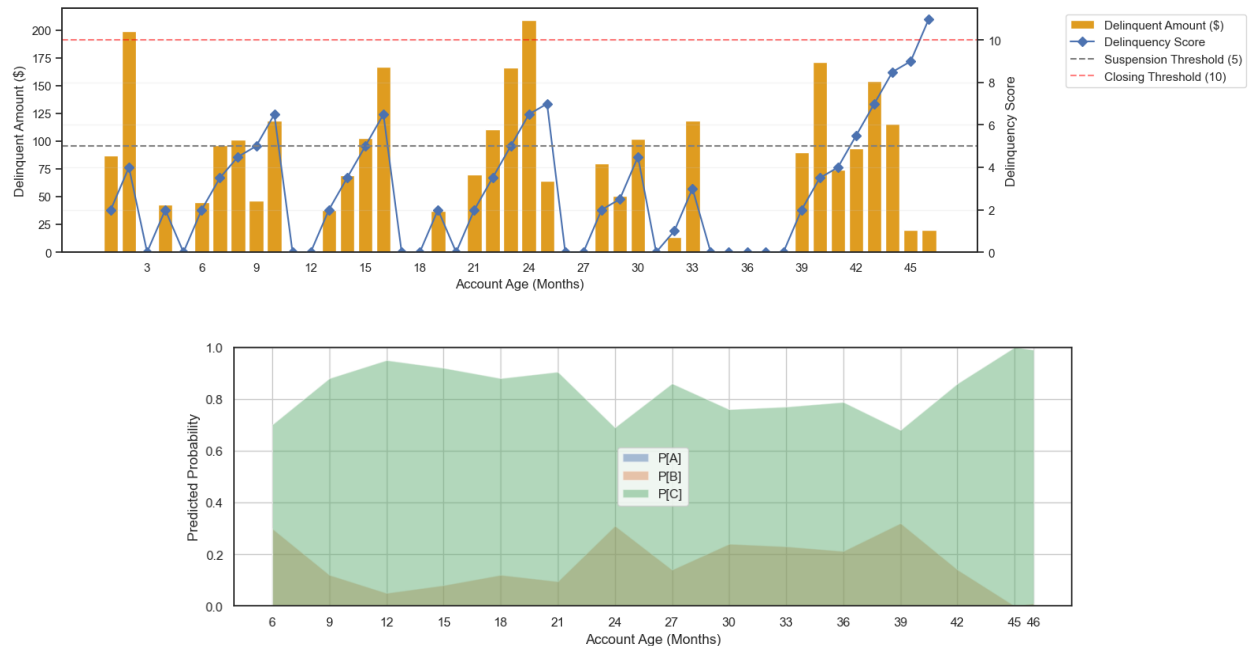
- Repeated missed or late payments, often leading to sharp spikes in delinquency score
- Several sustained penalty cycles, but not enough to trigger account closure
- A pattern of unstable behavior: problems weren't isolated, but also didn't fully escalate

### Interpretation:




- This is a watchlist-type account. It's not a clear default, but the pattern of repeated delinquency signals real concern. The model's classification reflects this uncertainty, placing the account firmly in the "at-risk but recoverable" category.

## Account C – High Risk

- **Status:** Closed
- **Account Duration:** 46 months
- **Delinquency History:** Persistent payment failures with multiple escalations—account crossed suspension thresholds repeatedly and eventually closed
- **Prediction Pattern:** High confidence in Class C throughout—probability >80% in most months after early signs of deterioration



### What the Model Saw:

-  Delinquency score steadily climbed, eventually breaching the closing threshold
-  Long and repeated streaks of missed payments
-  High penalty accumulation, no sustained recovery

### Interpretation:

- A clear high-risk case with structural payment issues and no lasting improvement. The model accurately identified the risk buildup and consistently flagged the account for intervention—well before closure.

## 9. Recap: From Behavior to Risk Classification

This phase of the project transformed raw billing data into a practical tool for predicting and understanding credit risk. Here's what we built—and why it matters.

### A Full Classification Pipeline

We developed an end-to-end system that:

- Aggregates monthly billing snapshots into lifecycle-aware features
- Labels accounts using a hybrid rule and clustering approach (Class A/B/C)
- Trains a supervised model to predict future risk based on past behavior
- Outputs class probabilities and explanation-ready predictions

## Features That Capture Behavior Over Time

We engineered features that track not just *what happened*, but *how it happened*:

- Missed payment rates and penalty severity
- Recovery streaks and recent issues
- Suspension patterns and long-term reliability

These metrics gave the model the ability to “understand” customer trajectories—critical for early detection.

## Labels That Reflect Business Reality

Instead of oversimplifying risk as a binary (good vs. bad), we introduced a three-class system:

- **Class A** – Consistently reliable
- **Class B** – Mixed or unstable
- **Class C** – High risk, closed or near-closure

This supports real-world use cases like triaging accounts, focusing collections, and tracking behavior shifts.

## Performance That’s Actionable

- Strong separation between classes in the confusion matrix
- High-confidence predictions aligned with clean behavioral signals
- SHAP values that explain *why* each account is classified the way it is

Together, these insights make the model not just accurate—but **usable by business teams**.

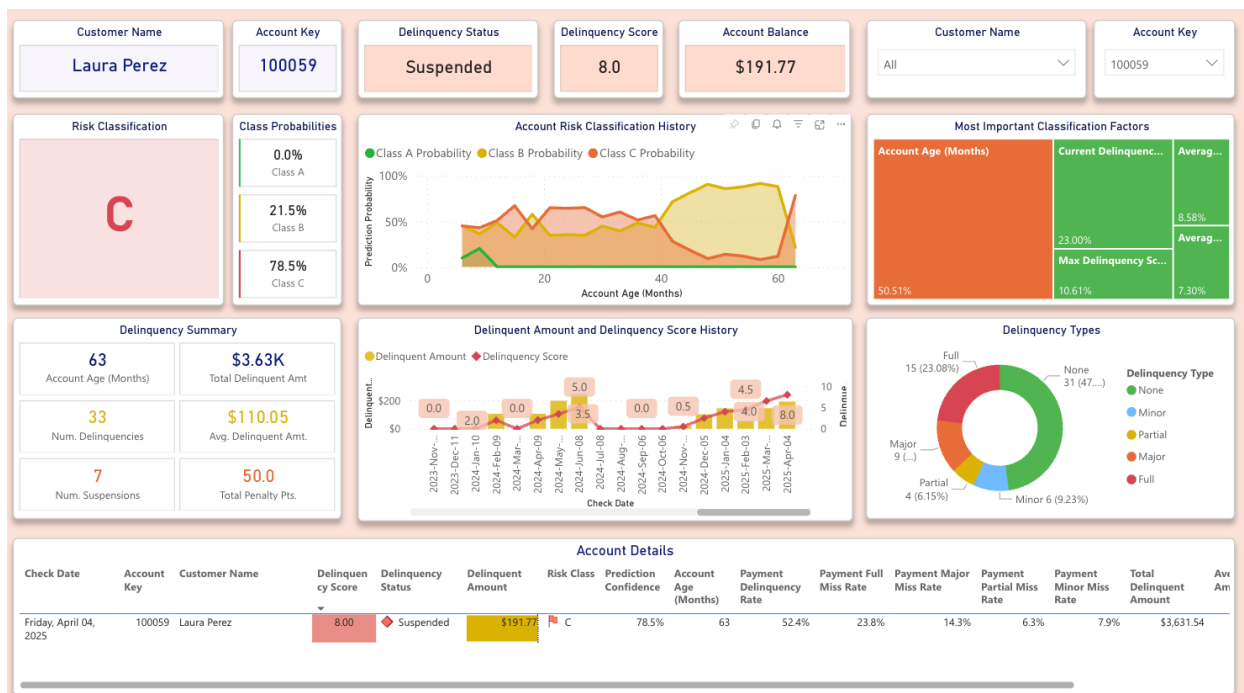
Next, in **Part 3**, we’ll bring this model to life in Power BI—turning predictions into real-time dashboards that help analysts monitor trends, drill into accounts, and take action with confidence.

---

# 10. Next Step: Bringing These Predictions to Life

## What’s Next?

In the final part of this series, we’ll bring everything together inside Power BI—showing how risk predictions become interactive dashboards that support daily operations, portfolio oversight, and explainable decision-making.



If you're interested in how to make machine learning useful—not just accurate—I invite you to explore **Part 3**, where we operationalize this model into a real-time credit risk monitoring tool.

## Linked Project Series: From Data Engineering to Scoring

This article is the second installment in a three-part series on building a full-cycle credit risk monitoring system for utility accounts. Each part builds on the last—creating a practical pipeline from behavioral Engineering to business-ready insight:

- Part 1 – Data Engineering**  
 We created a synthetic dataset that mirrors real utility billing behavior over time—tracking payments, penalties, suspensions, and closures. This simulation engine formed the foundation for everything that followed.
- Part 2 – Risk Classification Model (*This Article*)**  
 We engineered time-aware behavioral features, designed a hybrid labeling strategy, and trained a Random Forest model to classify accounts into risk tiers (A/B/C). We also built in explainability with SHAP values, enabling both predictive power and model transparency.
- Part 3 – Monitoring Dashboard in Power BI**  
 We turned these predictions into an interactive report that helps business users monitor portfolio-level risk, investigate specific accounts, and understand the behavioral drivers behind each risk score.

## Why It Matters

By combining data engineering, modeling, and visualization, this project delivers a complete, explainable risk solution—ready for real-world decision-making. Whether you're an analyst, product owner, or operations lead, the tools and logic presented here offer a replicable framework for data-driven credit risk management.



---

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.