Turning Invoices into Credit Signal

Multimodal AI to Predict Brazilian Invoice Defaults

CERC Team: Marcelo Maziero, Juliano Pereira, Kenji Noda, Anna Castro
MIT Advisors: Rodrigo Verdi, Caio Barbosa


Rose Dana


Lottie Schulte-Bockum

Problem Statement

CERC is Brazil's leading data infrastructure for banks and credit funds to register SMEs' (Small and Medium-sized Enterprises) invoices as receivables for credit access.

Invoice Flow

B2B Customer (Hotel) → +30d → SME (Bakery)

Walk-in Customer (Rose) → Acquirer (Worldline) → +30d → SME (Bakery)

SME (Bakery) ⇄ Bank (Itaú Bank) ⇄ Registry (CERC)

+Risk Scores

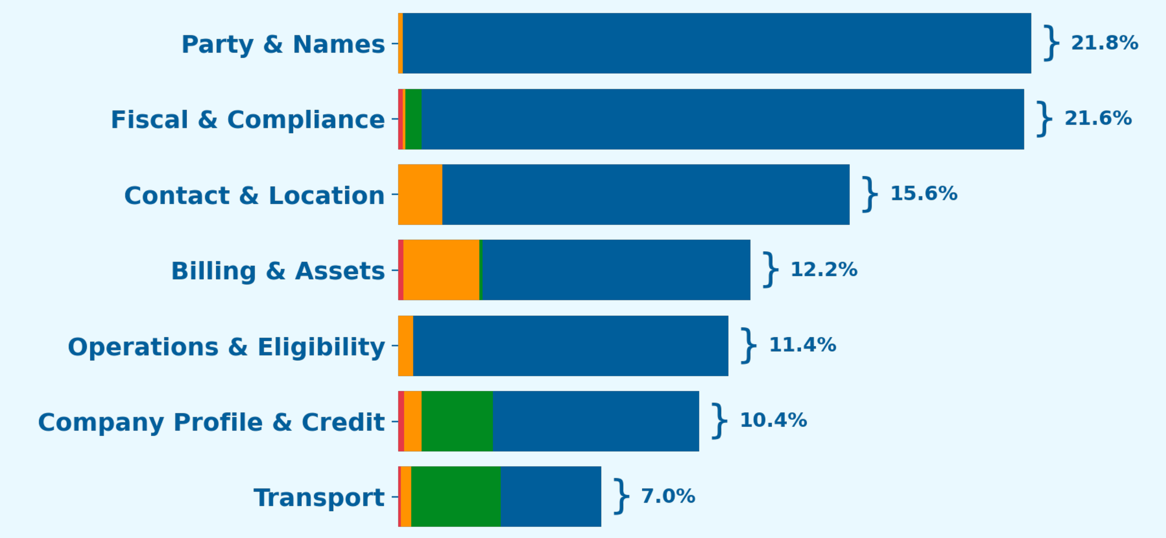
Problem: Banks currently rely on company-level heuristics to distribute credit but have no historical insights on invoice behavior. This means SMEs receive limited credit access at high interest rates.

Objective

Build Risk Scores: Leverage Invoice-level data that CERC assesses to produce default risk scores to help Banks score credit.

CERC's Invoice Assessments

CERC runs ~200 automated health & fraud checks for every registered invoice, verifying the parties, fiscal documents, transport events and billing data.



Each indicator returns a natural-language message, e.g.:

- 🇧🇷 Crítico → ***"CT-e cancelado"***
- 🇺🇸 Critical → ***"Transport e-document cancelled"***

Data Engineering

1. Credit Transactions

- Rev. past 7/30/60/90d
- Rev. Velocity
- Rev. Acceleration
- Rev. Stability
- Projected Rev.

2. Payer-Payee Network

Historic relationships between payers & payees

e.g. Closed Loop (A sells to B sells to C sells to A)

3. Indicator Response Embeddings

Concatenate Indicator messages (PT) → Encode MiniLM-L12-v2 *(384-dim)* → Reduce 384-dim → 32-dim via PCA

Methodology

No ground-truth default label exists, so we built two independent proxies

1. Engineered Features

1. Credit Transactions
2. Payer-Payee Network
3. Indicator Response Embeddings

2. Proxy Target Definition

Proxy 1: Repeated Relationship
When a new invoice is issued after the installment's due date

Invoice issued → Due date passed → New invoice

Proxy 2: Critical Indicator Change
When an invoice is reviewed at a later date and at least one indicator changes to Critical

Review 1 Neutro → Review 2 Crítico

3. Modeling Approach

INPUT
Engineered features + individual proxy

↓

TRAIN
Tree Ensembles
LightGBM · XGBoost · HistGradBoosting
Handles high-dim embeddings + NaNs

↓

OUTPUT
Default probability - per invoice, per proxy
Rank by probability

4. Proxy Verification

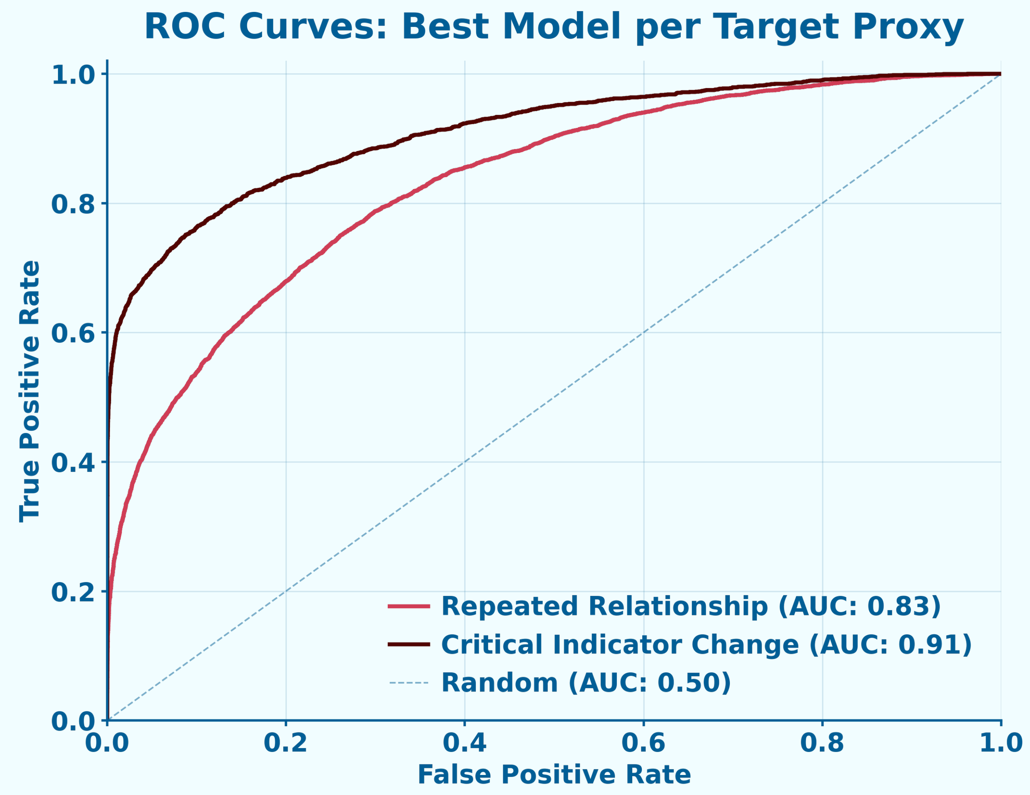
Spearman's ρ = Pearson correlation applied to the ranks of the two models' predictions.

Captures whether they order invoices by risk consistently, ignoring the raw probability scale.

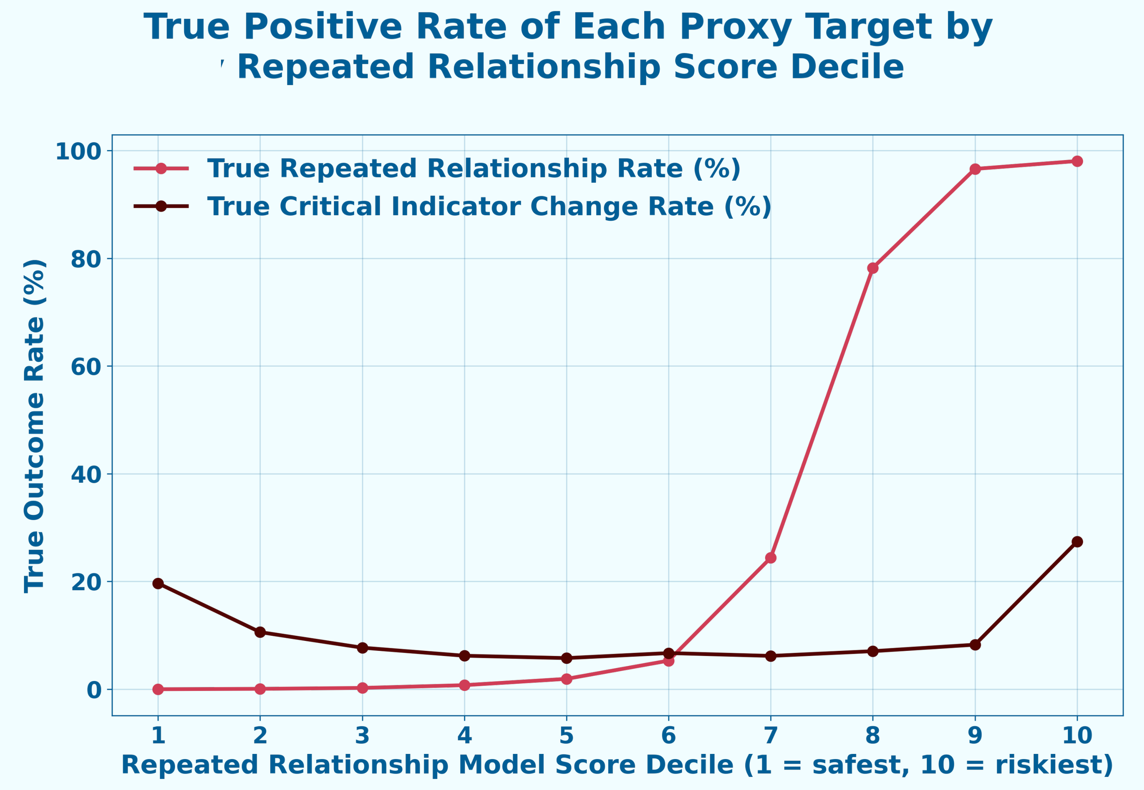
$$\rho_s = \frac{\sum_i (R(x_i) - \bar{R}_x)\,(R(y_i) - \bar{R}_y)}{\sqrt{\sum_i (R(x_i) - \bar{R}_x)^2 \cdot \sum_i (R(y_i) - \bar{R}_y)^2}}$$

x, y = invoice ranks under Proxy 1 & 2
$\bar{R}_x$, $\bar{R}_y$ = mean ranks
ρ ∈ [−1, 1], higher = stronger agreement

Results



Proxy	Best Model	AUC
Repeated Relationship	LightGBM	0.83
Critical Indicator Change	XGBoost	0.91



ρ_s = 0.202

Interpretation: The two proxies define default differently and flag largely different invoices, so the choice of proxy depends on the business context.

Business Impact

Market Risk		5% of CERC's R$250B (US$50B) invoice market defaults annually.
SMEs Impacted	**3.5M**	70% of Brazil's 5M SMEs are denied credit access. Using invoice signals will allow 3.5M SMEs to be considered.
Solution		Production-ready proxy-based ML Engine.
Value at Stake	**$1.7B**	With a recall of 0.694, our proxy-based model can catch up to R$8.7B (US$1.7B) annually.

CERC's Future

Duplicata Escritural: Regulation requiring all invoices to be registered digitally alongside payment information, bringing the true target variable to CERC in the long term.