

Predicting email-subscriber churn across 102 million sends

Análisis y modelización del riesgo de fuga en campañas masivas por email

Role: Data Analyst · Experian, Madrid · 2014 · **Origin:** MSc final thesis (TFM), Universidad Complutense de Madrid — Máster en Minería de Datos e Inteligencia de Negocios · Advisor: Dr. Javier Portela García-Miguel

CONTEXT

Email marketing platforms measured campaigns after the fact: opens, clicks, conversions — a description of what already happened. What the business could not answer was forward-looking: *which subscribers are about to go quiet, and what can we do about it?*

Voluntary unsubscribes are rare; silent inactivity is not. Modeling the frequent event rather than the rare one was the core reframing of the problem.

Data came from the Experian CheetahMail® / EMS platform for a retail fashion vertical: **1.42M unique subscribers** and **102.1M sends** over a 12-month observation window, with a 6-month look-back and a 4-month prediction horizon.

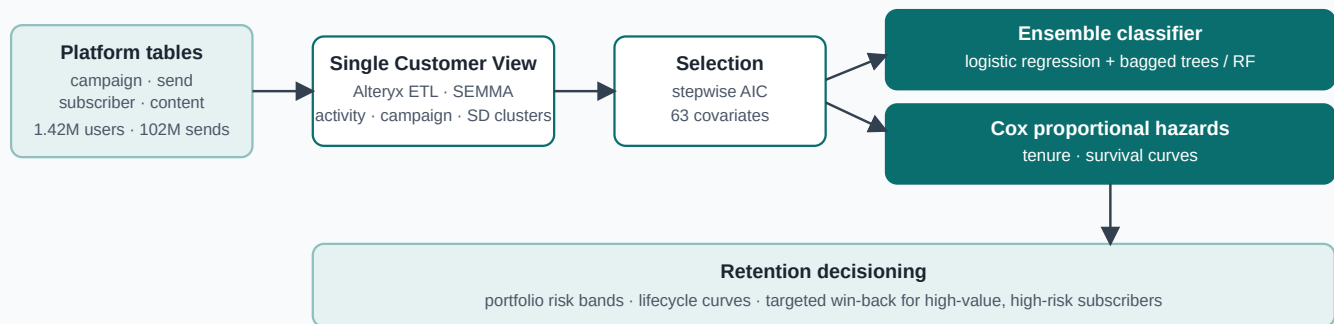
WHAT I BUILT

- A SEMMA pipeline over four platform tables (campaign, send, subscriber, content) collapsed into a Single Customer View — ETL in Alteryx, modeling in R.
- Engineered feature families: historic and recent activity (k-centroid clustering of engagement profiles), campaign typology and campaign-effectiveness clusters, tenure, and sociodemographic profile — reduced by stepwise AIC selection.
- An **ensemble classifier** averaging stepwise logistic regression with bagged decision trees / random forest, chosen because linear and partition-based learners misclassify different subscribers.
- A **Cox proportional-hazards model** on tenure to turn a binary prediction into a customer-lifecycle risk curve, and to band the whole portfolio by inactivity risk.

OUTCOME

- Ensemble reached **0.907 AUC** at a **15.6% misclassification rate** — against a 28.2% naive baseline, roughly halving the error.
- Cox model: concordance **0.868** across 63 covariates, producing survival curves and 0–25 / 25–50 / 50–75% portfolio risk bands finance and CRM could act on.
- Counter-intuitive finding that changed strategy: **send frequency and density were not significant drivers** of churn — how a subscriber *interacts* is. Saturating the list does not cause the exit; irrelevance does.
- Recent-activity variables reduced modeled inactivity risk by up to **22%**; voluntary disclosure of personal data (e.g. a mobile number) predicted materially longer lifecycles — an early trust-as-engagement signal.
- The approach was later integrated into Experian's CheetahMail® retention services.

MODELING PIPELINE



Validation: train/test partition on 198,124 subscribers (30% inactive within 4 months); model chosen on test misclassification, ties broken on ROC AUC.

0.907

ROC AUC, ensemble classifier

15.6%

misclassification, vs 28.2% baseline

102M

email sends modeled, 1.42M subscribers

0.868

Cox concordance on tenure

STACK

R Alteryx Logistic regression Bagging / Random Forest Ensemble methods Cox proportional hazards Survival analysis
k-centroid clustering SEMMA SQL

Based on my own MSc thesis, of which I hold full rights. Describes methodology and results only — no subscriber records, client identities or proprietary platform internals are disclosed. Experian and CheetahMail are trademarks of Experian Information Solutions, Inc., referenced here solely to identify where the work was carried out.