ANALYTICAL REPORT

Transformer Failure Risk Prediction

Self-directed portfolio analysis on a public utility dataset

Classification | Logistic Regression | Naive Bayes | Random Forest

Prepared for	Arbutus Visual Analytics
Prepared by	Marc Charbonneau, Principal Analyst, Arbutus Visual Analytics
Date	July 7, 2026
Confidentiality	Self-directed portfolio piece: public dataset, no client

EXECUTIVE SUMMARY

Written for a non-technical audience.

This report builds and compares three statistical models to estimate which electrical distribution transformers are at elevated risk of failure, using data a utility already collects; the strongest of the three, a Random Forest model, correctly flags 43% of transformers that fail within a year while keeping false alarms to about 4% of healthy ones. See Conclusions, right after the infographic overview, for the full findings and recommendations.

TRANSFORMERS ANALYZED
15,873
Distribution transformers in the 2019 dataset used to build and test the model
↓ *jump to finding*

HISTORICAL FAILURE RATE
5.08%
Share of the 15,873 transformers that failed ('burned') during 2019
↓ *jump to finding*

FAILURES CORRECTLY FLAGGED (RECALL)
42.9%
Random Forest, the best of three models tested, caught about 43 of every 100 transformers that actually failed. A different, harder measure than overall accuracy (93.7%); below the one verified same-dataset benchmark (52-56%, Section 4.4), a gap now narrowed to 9-13 points and traced to one remaining disclosed methodology difference
↓ *jump to finding*

MODEL FIT SCORE (F1)
0.409
A single 0-to-1 score balancing catching real failures against false alarms; higher is better. Random Forest's score, the best of the three models compared on equal footing
↓ *jump to finding*

RANKING ABILITY (ROC-AUC)
0.851
How well Random Forest, the best-performing model, ranks at-risk transformers above healthy ones, from 0.5 (random guessing) to 1.0 (perfect)
↓ *jump to finding*

ASSUMPTION-VIOLATION PENALTY (NAIVE BAYES F1)
0.188
Naive Bayes, built specifically to illustrate an independence-assumption violation, scored 0.188 vs. logistic regression's 0.326 on an identical feature set (Section 4.3); a separate illustration from Random Forest's stronger performance above, not a competing 'best model' claim
↓ *jump to finding*

ENGAGEMENT OVERVIEW

Predicting Electrical Transformer Failure Risk

A public-data flagship analysis for Arbutus Visual Analytics

1 PROBLEM

The Challenge

- Which transformers in a distribution network are more likely to fail within a year?
- Can load, lightning exposure, and network characteristics predict that risk before failure happens?
- How do three different statistical approaches, including a nonlinear ensemble model, compare on the same imbalanced, real-world dataset?

2 WORKFLOW

How We Built It

- Cleaned and audited 15,873 transformer records across 16 recorded characteristics
- Explored distributions, class separation, and correlations before choosing a model
- Engineered and encoded features to suit each model type
- Built logistic regression and Naive Bayes on an identical feature set to illustrate an assumption violation, then Random Forest on the full predictor set to establish the strongest result achievable
- Tuned each model's decision threshold using training data only, then tested all three on data held back until the end

3 SOLUTION

What We Delivered

- A risk-scoring model (Random Forest) that flags 43% of true failures while keeping false alarms to about 4%
- A clear, interpretable set of risk drivers from a companion logistic regression model: lightning exposure, network length, and equipment characteristics
- A side-by-side comparison showing how an assumption violation quietly degrades a simpler model
- A narrowed, still-documented gap against published literature, with one remaining concrete next step identified for further improvement

Dataset: 'Dataset of Distribution Transformers at Cauca Department (Colombia)' (Bravo Montenegro, Alvarez, Lozano; Mendeley Data, CC BY 4.0)

Conclusions

Written for a non-technical audience.

WHAT WE SET OUT TO ANSWER

Can statistical models reliably identify which electrical distribution transformers are more likely to fail within a year, using only load, lightning exposure, and network characteristics a utility already records?

WHAT WE FOUND

- A Random Forest model, the best-performing of three models built, correctly identified 43% of transformers that failed within the year, while keeping false alarms to about 4% of healthy transformers, using only data utilities already collect.
- A companion logistic regression model, though less accurate (41% recall), showed directly and interpretably that longer power line networks and higher local lightning activity are the two strongest, most actionable risk signals identified.
- A third model built specifically for comparison (Naive Bayes) performed markedly worse once several related risk factors were added together, a concrete demonstration of why a model's assumptions need to be checked, not just its final accuracy number.
- A published study on this exact dataset, using a nonlinear model and a rebalanced training set, still achieves a meaningfully higher failure-detection rate (52-56% versus Random Forest's 43%); adding a nonlinear model family narrowed that gap from 11-15 points to 9-13 points, leaving one clear, achievable next step: resampling the training data.
- The dataset comes from a Colombian utility and is used here as a public stand-in for utility asset risk modelling methodology, not as a finding about any Canadian utility's actual equipment.

WHAT WE RECOMMEND

1. Use the Random Forest model to prioritize physical inspections toward the small subset of transformers it flags as highest-risk, rather than inspecting on a fixed schedule alone.
2. Track additional data most utilities do not currently capture, such as localized surge events and harmonic distortion, since these are known real-world failure drivers this dataset could not test.
3. Treat the simpler comparison model (Naive Bayes) as a cautionary example rather than an operational tool, unless its known statistical assumption violations are first addressed.

Abbreviations

AUC	Area Under the (ROC) Curve
EENS	Electric Energy Not Supplied, in kilowatt-hours
F1	F1 score: the harmonic mean of precision and recall
kVA	Kilovolt-Ampere, a unit of transformer rated capacity
LR	Logistic Regression
NB	Naive Bayes
RF	Random Forest
ROC	Receiver Operating Characteristic
SD	Standard Deviation

1. Data Source, Cleaning, and Variable Roles

> This section covers the dataset's structure and provenance, the cleaning checklist applied before any modelling, and how each of the sixteen recorded characteristics was assigned a role in the model.

1.1. Dataset and Study Population

The dataset covers 15,873 electric power distribution transformers operated by Compania Energetica de Occidente in the Cauca department of Colombia, observed across 2019 and 2020 (Bravo Montenegro, Alvarez, and Lozano, *Data in Brief*, 2021; CC BY 4.0). Sixteen columns are recorded per transformer per year, covering load, lightning exposure, network characteristics, connection type, and a binary annual failure ('burned') flag. Direct inspection of both source files confirmed the two years describe the same 15,873 physical transformers rather than two independent samples: 11 of 13 static predictor columns are identical between years, and of the 807 transformers that burned in 2019, only 24 (about 3%) burned again in 2020. Modelling both years as 31,746 independent rows would therefore pseudo-replicate near-identical predictor vectors and violate the row-independence assumption most classifiers rely on. The 2019 cross-section (15,873 rows, 807 failures, a 5.08% failure rate) was used as the sole modelling dataset; 2020 was held out entirely as a candidate out-of-time robustness check for future work.

1.2. Cleaning

Both source files loaded cleanly with no header issues, no non-UTF-8 characters, and no placeholder strings in place of numeric values. Every column already carried the correct data type on load, and zero missing values were found in any column, in either year. Seventy-eight exact duplicate rows (about 0.5% of 2019) were identified and kept: the dataset has no unique transformer identifier, several of its columns are coarse and low-cardinality (two binary flags, two small categorical fields), and legitimate collisions between distinct, small, standard-sized transformers are expected at this sample size. One column, the length of low-tension secondary line per transformer, showed a smooth but heavily right-skewed distribution (70th percentile 14,605 metres, 90th percentile 159,722 metres, maximum 717,125 metres) consistent with genuinely long rural feeders rather than a data-entry error; it was log1p-transformed rather than capped or dropped, since log1p is defined at the exact-zero values present in 1,008 rows. The same log1p transform was later applied to three other right-skewed continuous predictors, rated capacity, customer count, and energy not supplied, once the model family was chosen (Section 3).

1.3. Variable Roles

The dataset contains no unique transformer identifier, so no identifier-exclusion decision was required. Ten of the sixteen raw columns entered the model as predictors: five continuous (rated transformer capacity in kVA, average and maximum zone-level lightning discharge density in rays per square kilometre per year, a historical burning rate, customer count, EENS, and network length), five binary flags (urban/rural location, self-protection, ceramic-insulator criticality, removable connectors, and circuit queue), and two nominal categorical fields (installation type and client type). One near-constant binary flag, indicating whether a transformer is connected via an aerial network, was excluded: only 33 of 15,873 rows (0.2%) took the minority value, and every one of those 33 rows had a negative failure outcome, leaving nothing for a model to learn beyond memorising 33 specific rows. Installation type's smallest categories (a single cabinet-type transformer, three metal-tower units, 32 pad-mounted units, and 56 units in a residual 'other' group, 92 rows combined) were consolidated into a single 'OTHER' level before encoding, to avoid one-hot columns with too few observations to learn from; the four categories with meaningful sample sizes (pole-mounted, macro without anti-fraud net, pole with anti-fraud net, and H-type) were kept as their own levels. Client type's nine levels, six of which correspond to Colombia's residential socioeconomic stratification (Stratum 1 through 6) and three of which are non-residential categories (commercial, industrial, and official) with no principled position on that same scale, were treated as a single nominal field and one-hot encoded in full, since a linear model requires nominal categories to be one-hot rather than label-encoded; the stratum ordering is instead discussed qualitatively wherever coefficients for those categories are interpreted.

One predictor's directionality could not be fully confirmed against the source literature. The location field is a binary urban/rural indicator, per the primary source paper's own variable

dictionary, but a secondary paper reusing this dataset reports counts that appear to reverse which numeral (0 or 1) corresponds to 'urban' relative to the primary paper's stated definition. Given that conflict, the model uses location as an unordered binary flag without asserting a specific urban-versus-rural direction for either value; this is disclosed as a limitation rather than resolved by assumption.

1.4. A Data Point That Looked Like Leakage but Was Not

EENS is defined in the primary source paper as 'the kilowatt-hours the distribution company stops selling when the transformer stops operating due to a failure event,' wording that, read in isolation, sounds like a same-year consequence of the outcome being predicted: a data leakage risk. This was checked directly rather than assumed: transformers that burned in 2019 actually show a lower mean and median EENS value (328 kWh mean, 243 kWh median) than transformers that did not burn (558 kWh mean, 405 kWh median), the opposite pattern a same-transformer, same-year post-outcome calculation would produce, and consistent with the variable's weak negative correlation with the failure outcome ($r = -0.063$). This indicates EENS behaves as a pre-existing systemic or circuit-level risk index, similar to the historical burning rate, rather than a literal after-the-fact consequence of that specific transformer's own failure. It was retained as a predictor on this basis.

2. Exploratory Data Analysis

> Distributions, class separation, multivariate relationships, correlation structure, and outlier screening were reviewed before any model was fit.

2.1. Distributions

Figure 1 shows right-skewed distributions for rated capacity, customer count, and EENS, consistent with a mostly small, residential-scale transformer fleet with a thin tail of large industrial units. The two lightning-density variables and the historical burning rate show discrete, multi-modal spikes rather than smooth curves: lightning density is recorded at the geographic-zone level, so every transformer in the same zone shares an identical value, and the burning rate is a coarse historical rate metric (0, 0.25, 0.5, and so on) rather than a continuously measured quantity. Network length is bimodal, with a spike at zero (1,008 transformers with no secondary line recorded) and a second cluster corresponding to longer rural feeders; this shape is what motivated the log1p transform described in Section 1.2 rather than a plain log transform, since log1p is defined at zero.

Figure 1. Distribution of continuous predictors, split by failure status.

2.2. Separation Between Failed and Healthy Transformers

Figure 2 splits each continuous predictor by failure outcome. Both lightning-density variables show a visibly higher median and interquartile range for failed transformers, consistent with lightning exposure contributing to failure risk. Network length shows a clearly higher median for failed transformers (about 11.7 on the log1p scale versus about 10.5 for healthy transformers), consistent with longer rural feeders carrying more risk. The historical burning rate's upper quartile is higher for failed transformers despite both groups sharing a median of zero. Rated capacity, customer count, and EENS all show a narrower range and fewer large values among failed transformers than healthy ones, the reverse of a naive 'bigger transformer, bigger risk' intuition, and a pattern later reflected in the model's results (Section 4).

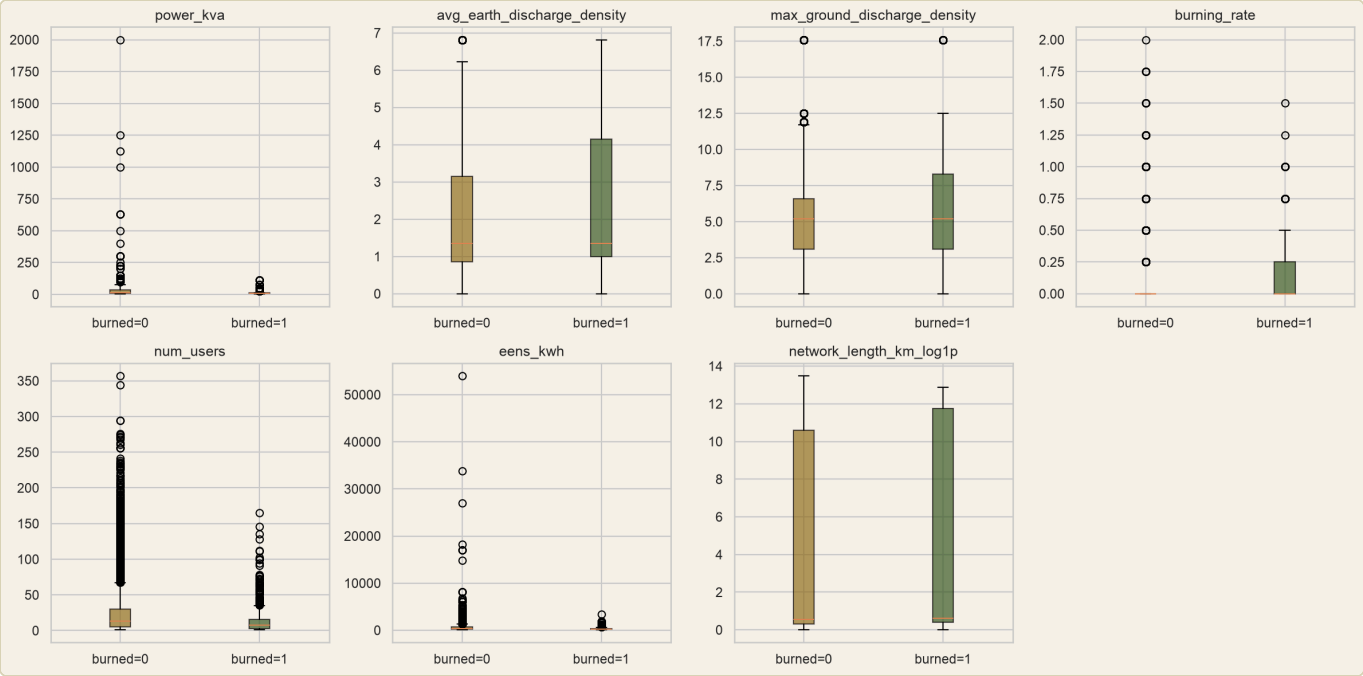

Figure 2. Box plots of continuous predictors, split by failure status.

2.3. Multivariate Relationships

Figure 3 (a 2,000-row subsample) shows the same relationships jointly rather than one variable at a time. Failed transformers, shown in a separate colour, appear throughout the predictor space rather than in one isolated region, consistent with failure risk depending on a combination of factors rather than any single variable acting as a clean dividing line. This figure exists to check for that kind of clean separation before committing to a linear model; its absence here supports treating this as a genuinely multivariate classification problem rather than one solvable by a single strong predictor.

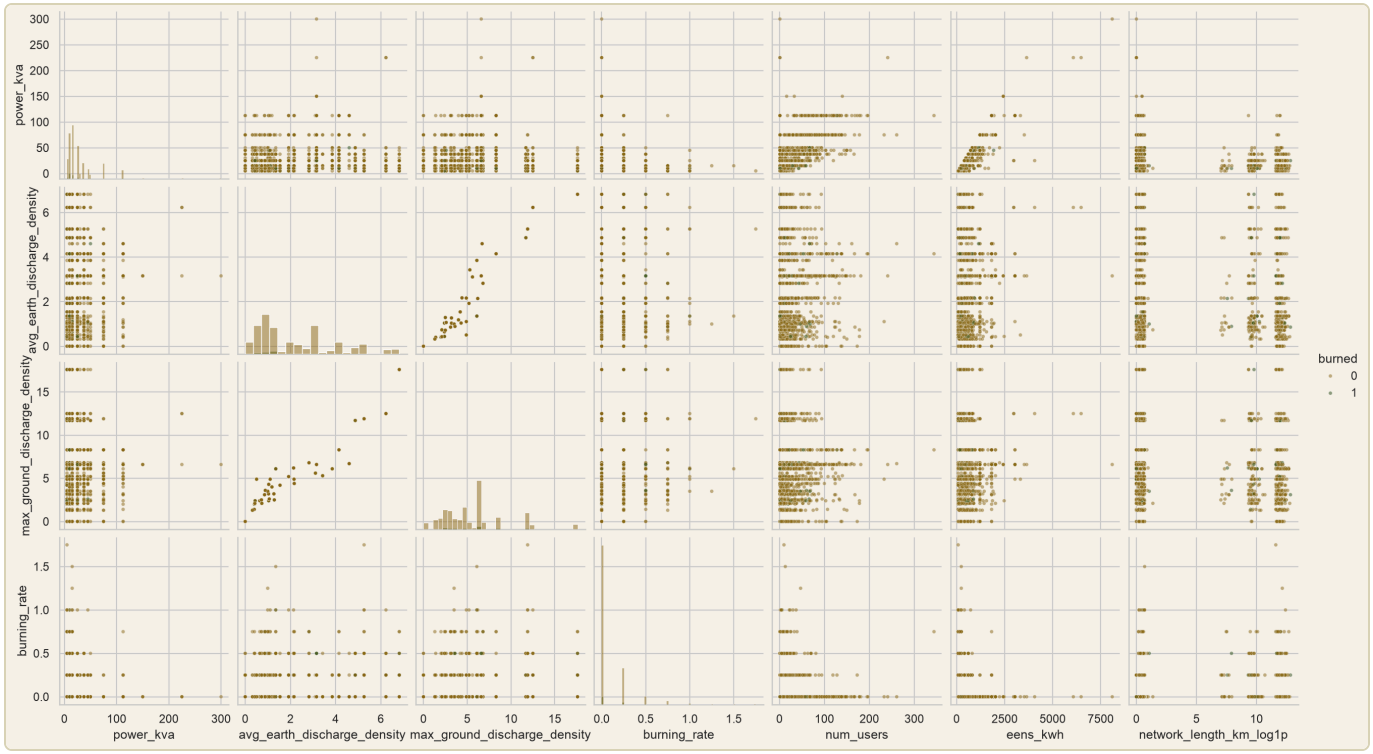

Figure 3. Scatterplot matrix of continuous predictors, split by failure status (2,000-row subsample; 1 of 2).

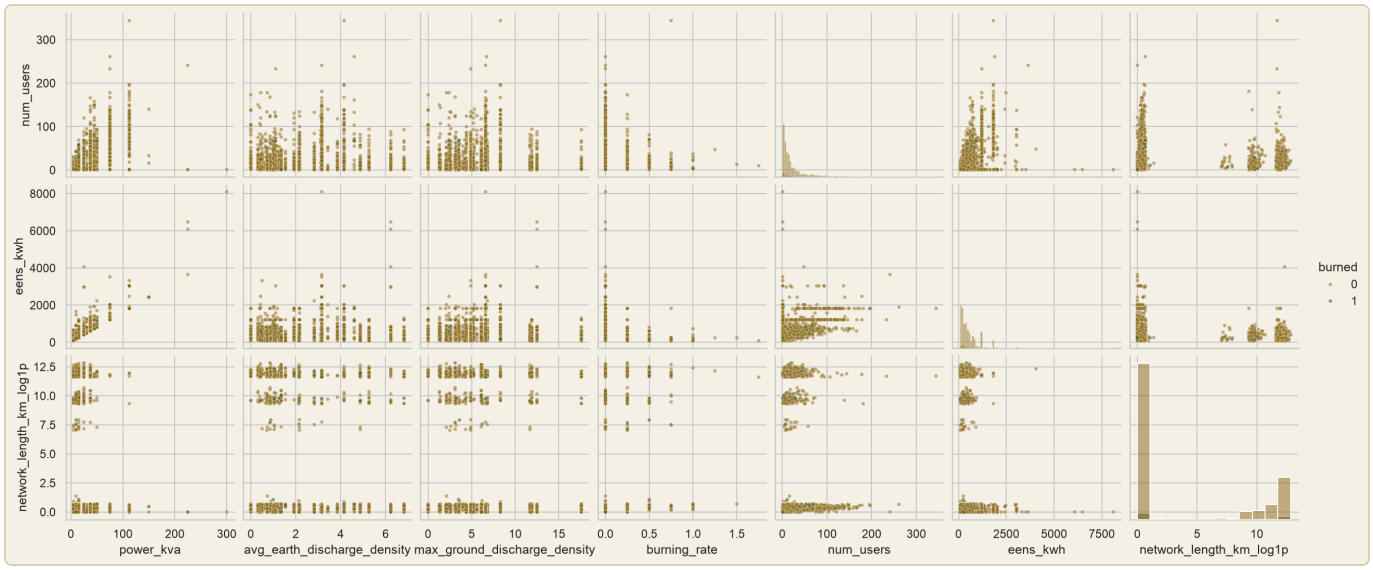

Figure 3 (continued). Scatterplot matrix of continuous predictors, split by failure status (2,000-row subsample; 2 of 2).

2.4. Correlation Structure

Figure 4 flags two near-duplicate predictor pairs at or above a correlation of 0.5, the threshold at which two predictors start to make individual logistic regression coefficients hard to interpret independently: rated capacity and EENS (r = 0.94), and average and maximum lightning discharge density (r = 0.94). A third pair, rated capacity and customer count (r = 0.53), sits just above that same threshold. Both near-duplicate pairs are two different ways of summarising essentially the same underlying quantity (kVA rating tracks typical energy throughput; average and maximum lightning

density are two summaries of the same zone climatology), and both are resolved before modelling (Section 3.2).

Figure 4. Correlation matrix, continuous predictors and the failure outcome.

2.5. Outlier Screening

Observations more than three standard deviations from the mean were flagged, not removed, for five continuous predictors: maximum lightning discharge density (402 rows), historical burning rate (315 rows), customer count (379 rows), EENS (111 rows), and rated capacity (58 rows). All five are consistent with the heavy right-skew already visible in Figure 1 rather than new data-quality concerns, and none were removed prior to modelling.

3. Model Design and Methodology

> Three model families, logistic regression, Naive Bayes, and Random Forest, were built and compared using consistent encoding, scaling, imbalance-handling, and threshold-tuning procedures across each.

3.1. Response Variable and Model Families Compared

The outcome, whether a transformer failed ('burned') during 2019, is binary, with a 5.08% positive rate. No model family was specified in advance for this self-directed piece. Two model families were built first and compared directly: logistic regression, chosen for its interpretable coefficients, and Gaussian Naive Bayes, chosen specifically to demonstrate, empirically, the practical cost of violating its own feature-independence assumption (Section 3.2 and Section 4.3). A third family, Random Forest, was added afterward: an initial benchmark comparison against a published study on this exact dataset (Section 4.4) showed both linear/probabilistic models underperforming that benchmark by 11-15 recall points, and Random Forest was chosen as a best-effort nonlinear alternative because it handles the correlated predictor pairs (Section 3.2) and any non-linear relationships natively, requires no feature scaling or resampling, and needed minimal additional tuning to add to this comparison. All ten predictors confirmed in Section 1.3 formed the starting candidate set for every model; an incremental cross-validation process (Section 4.1) added them in blocks to justify which blocks earned a place in each model's final feature set, rather than excluding any predictor upfront on correlation-with-outcome grounds alone.

3.2. Multicollinearity and the Naive Bayes Independence Assumption

The two near-duplicate predictor pairs flagged in Section 2.4 required different remedies for logistic regression and Naive Bayes. For logistic regression, one predictor from each pair was dropped: rated capacity was dropped in favour of EENS (the two are 0.94-correlated, and EENS has a more direct risk narrative for a utility audience), and average lightning discharge density was dropped in favour of maximum discharge density (which had the stronger raw correlation with the failure outcome: 0.042 versus 0.017). For Naive Bayes, within-class correlations, computed separately for failed and healthy transformers since Naive Bayes assumes features are independent conditional on the outcome class, confirmed both dropped pairs were resolved by the same two exclusions, but surfaced two further violations of the model's 0.3 independence threshold that persisted after those drops: customer count and EENS ($r = 0.52$ among failed transformers, 0.46 among healthy transformers), and historical burning rate and EENS ($r = -0.44$ and -0.34 respectively). Naive Bayes was kept on the identical feature set as logistic regression rather than dropping further predictors for Naive Bayes alone, to preserve a clean, directly comparable side-by-side result; the cost of that choice is demonstrated empirically in Section 4.3.

3.3. Random Forest's Predictor Set

Random Forest is not a linear model: correlated predictors do not destabilise its splits or inflate coefficient variance the way they do for logistic regression, so the multicollinearity remedy in Section 3.2 (dropping one predictor from each near-duplicate pair) solves a problem Random Forest does not have. Applying that same reduced, eight-predictor feature set to Random Forest anyway would understate its achievable performance without a corresponding methodological reason. Random Forest was therefore built on the full ten-predictor candidate set confirmed in Section 1.3, restoring rated capacity and average lightning discharge density; logistic regression and Naive Bayes remain on the reduced, eight-predictor set throughout this report. This is the one place the three models' inputs differ, and it is a deliberate choice rather than an inconsistency.

3.4. Gaussian Assumption, Class Imbalance, and Scaling

Gaussian Naive Bayes additionally assumes each continuous predictor is approximately normally distributed within each outcome class. Checked directly on the final five continuous predictors, the historical burning rate was found severely non-Gaussian in both classes (75-77% exact zeros, skewness above 2.5), a zero-inflated shape no transform corrects; it was kept regardless, since dropping a domain-relevant predictor from only one of the two models would break the like-for-like comparison, and the violation is disclosed here as a limitation. The remaining four continuous predictors were reasonably close to Gaussian after their log1p transforms. Given the 5.08% failure rate, all three models were fit with class-balanced weighting (a balanced class-weight setting for logistic regression and Random Forest; the equivalent balanced sample-weighting for Naive Bayes) rather than synthetic oversampling, keeping every model adjusted for the same imbalance through its own native weighting mechanism instead of favouring one with resampled training data. Continuous predictors were standardised (zero mean, unit variance, fit on the training set only) for logistic regression; Naive Bayes and Random Forest used unscaled continuous predictors, since both are invariant to linear rescaling (Naive Bayes by its per-feature probability model, Random Forest because tree splits do not depend on a feature's numeric scale). One-hot encoded categorical and binary columns were left unscaled throughout. Random Forest's hyperparameters (300 trees, a minimum of 5 samples per leaf to limit overfitting given only 646 positive training rows) are reasonable defaults, not tuned via grid or random search; a tuned Random Forest would likely score higher still, and this is disclosed as a limitation rather than presented as an exhaustively optimised ceiling.

3.5. Train/Test Split and Threshold Tuning

The 15,873 2019 transformers were split 80/20 (12,698 training rows, 646 failures; 3,175 test rows, 161 failures), stratified on the failure outcome so both sides preserved the same 5.08% failure rate, with a fixed random seed (42) for reproducibility. F1 score was chosen as the primary evaluation metric ahead of any model being trained, since no client-specific cost figures exist to justify optimising for recall or precision alone. The default 0.5

classification threshold rarely maximises F1 for a class-weight-balanced model on a 5.08%-positive outcome, so each model's decision threshold was tuned to maximise F1 using out-of-fold predictions on the training set only, then applied unchanged to the held-out test set, avoiding any test-set leakage into threshold selection.

4. Results and Model Comparison

> All three models were compared through an incremental feature-block cross-validation, then evaluated once on the held-out test set; a post-hoc comparison against published literature on this dataset followed and motivated adding the third model.

4.1. Incremental Feature-Block Comparison

Predictors were added in five cumulative blocks, with five-fold cross-validated F1 scores reported for all three models at each step (Table 1). Logistic regression and Naive Bayes share the same cumulative predictor counts (reduced, eight-predictor set); Random Forest's counts run one or two higher at each step, since it carries the two additional predictors restored in Section 3.3.

FEATURE BLOCK	PREDICTORS (LR/NB, RF)	LR F1 (MEAN, SD)	NB F1 (MEAN, SD)	RF F1 (MEAN, SD)
Lightning exposure	1, 2	0.098 (0.011)	0.109 (0.020)	0.130 (0.007)
+ Risk and load history	4, 6	0.141 (0.006)	0.143 (0.004)	0.209 (0.012)
+ Network and geography	6, 8	0.139 (0.003)	0.124 (0.003)	0.210 (0.013)
+ Equipment flags	10, 12	0.274 (0.019)	0.176 (0.012)	0.345 (0.012)
+ Categorical (installation, client type)	22, 24	0.282 (0.013)	0.115 (0.001)	0.338 (0.017)

Logistic regression's F1 score rose steadily as predictor blocks were added, from 0.098 with lightning exposure alone to 0.282 with the full feature set, with the equipment-flags block (0.141 to 0.274) contributing the largest single gain. Naive Bayes improved through the first three blocks but then declined once the equipment-flags and categorical blocks were added (0.143 down to 0.115), the empirical signature of the independence-assumption violations logged in Section 3.2: adding correlated predictors caused Naive Bayes to double-count shared evidence rather than add genuinely new information. Both models were kept on the identical, full reduced feature set regardless of this divergence, specifically to present this comparison as a direct, worked illustration of the cost of violating Naive Bayes' independence assumption, rather than quietly optimising Naive Bayes' own feature set to hide

it. Random Forest led every block from the first (0.130 versus logistic regression's 0.098) and, unlike Naive Bayes, never declined as predictors were added, rising to 0.345 at the equipment-flags block before easing slightly to 0.338 with the categorical block added, consistent with a model family that is not destabilised by correlated or lower-value predictors the way Naive Bayes is.

4.2. Final Test-Set Performance

Table 2 reports accuracy, precision, recall, F1, and ROC-AUC for all three models on the 3,175-row held-out test set, at each model's own training-tuned threshold.

METRIC	LOGISTIC REGRESSION	NAIVE BAYES	RANDOM FOREST
Accuracy	91.4%	93.2%	93.7%
Precision	27.0%	23.8%	39.2%
Recall	41.0%	15.5%	42.9%
F1 score	0.326	0.188	0.409
ROC-AUC	0.806	0.761	0.851
True positives	66	25	69
False positives	178	80	107
False negatives	95	136	92
True negatives	2,836	2,934	2,907

Accuracy alone is a misleading headline number for this dataset: with only 5.08% of transformers failing, a model that predicted 'never fails' for every single transformer would score 94.9% accuracy while catching zero real failures. Recall (the share of actual failures the model catches) is the metric that measures whether a model is actually useful here, and is a materially harder number to move than accuracy. On that basis, Random Forest is the strongest of the three models on every reported metric: an ROC-AUC of 0.851, an F1 score of 0.409, and a recall of 42.9% (69 of 161 actual failures correctly flagged) at a 3.6% false-positive rate (107 of 3,014 healthy test transformers), and is the model recommended for any operational use of these results. Logistic regression reached a close second on recall (41.0%, 66 of 161) but a noticeably lower F1 (0.326) and ROC-AUC (0.806), reflecting weaker precision (27.0% versus Random Forest's 39.2%); its coefficients remain the more directly interpretable of the two, which is why Section 3.1 kept it in the comparison rather than replacing it outright. Naive Bayes trails both on every metric (15.5% recall, 0.188 F1), and its higher accuracy (93.2%) paired with its much lower recall is itself a further

illustration of why accuracy alone should not be trusted on an imbalanced outcome. Figure 5 shows all three models' ROC curves. Figure 6 and Figure 7 show the logistic regression and Random Forest confusion matrices side by side, the two models with genuine operational recall; Figure 8 shows the Naive Bayes confusion matrix separately, since it is presented here as a cautionary counterexample rather than an operational candidate.

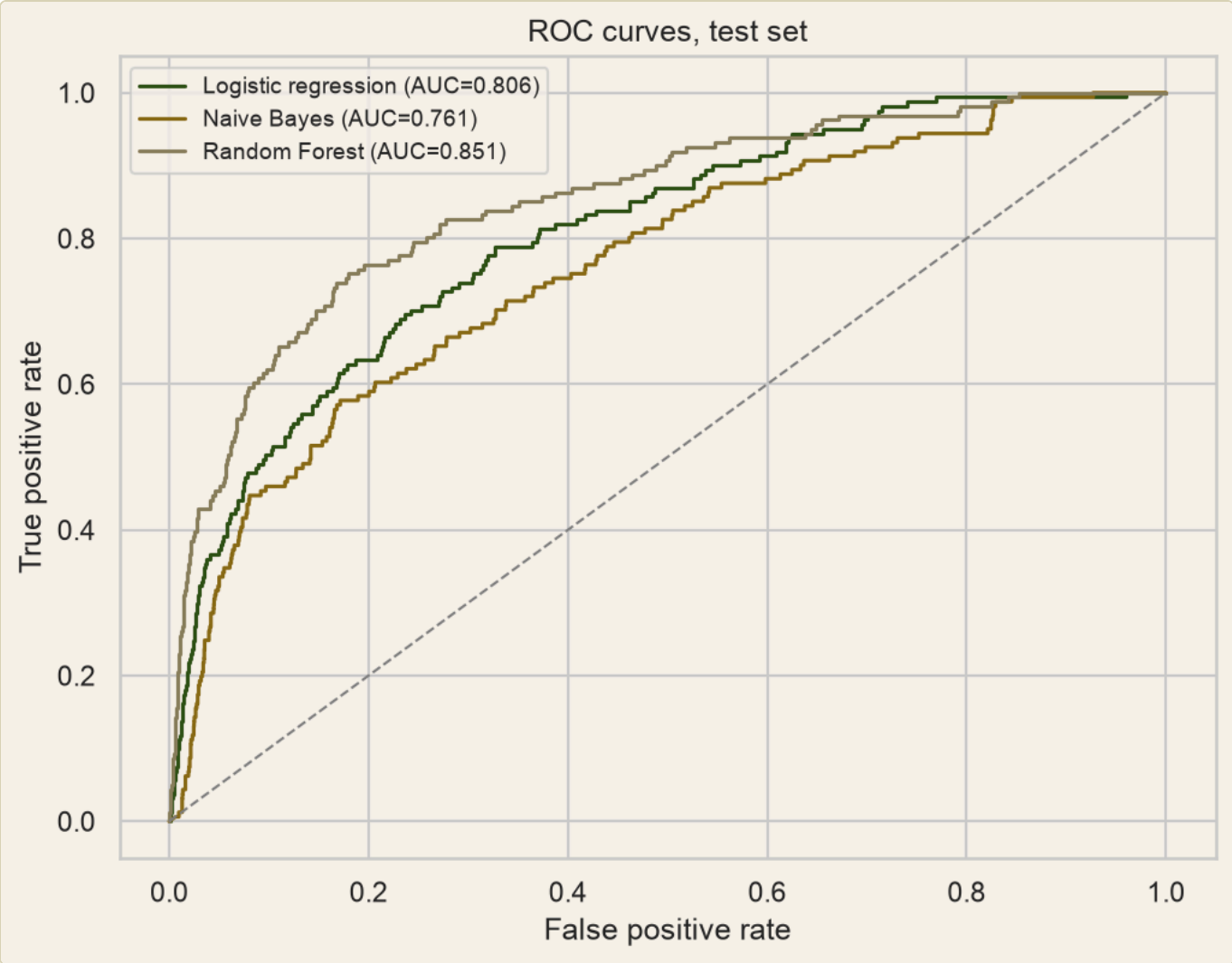

Figure 5. ROC curves for all three models, test set.

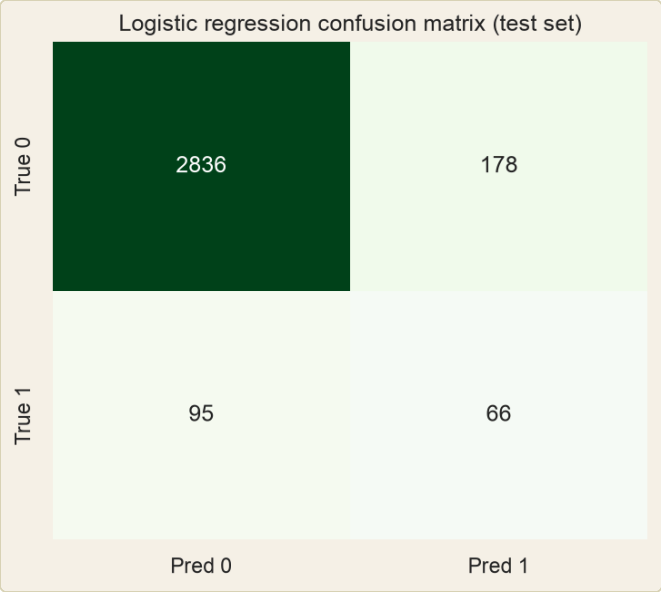

Figure 6. Logistic regression confusion matrix, test set.

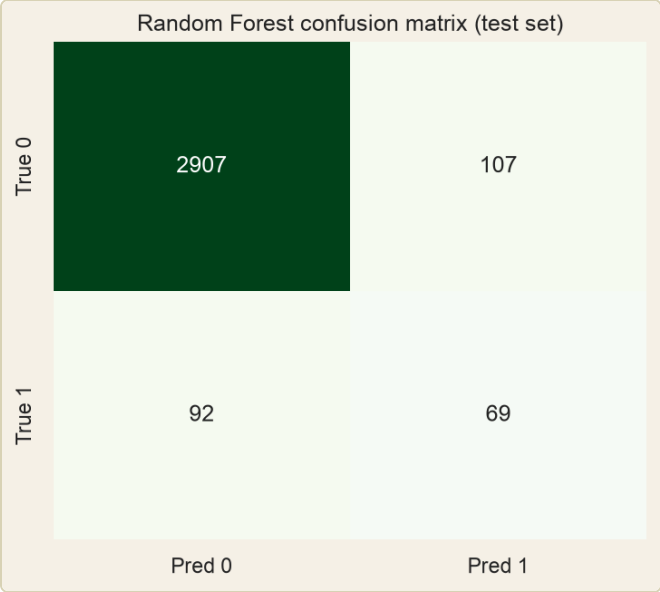

Figure 7. Random Forest confusion matrix, test set.

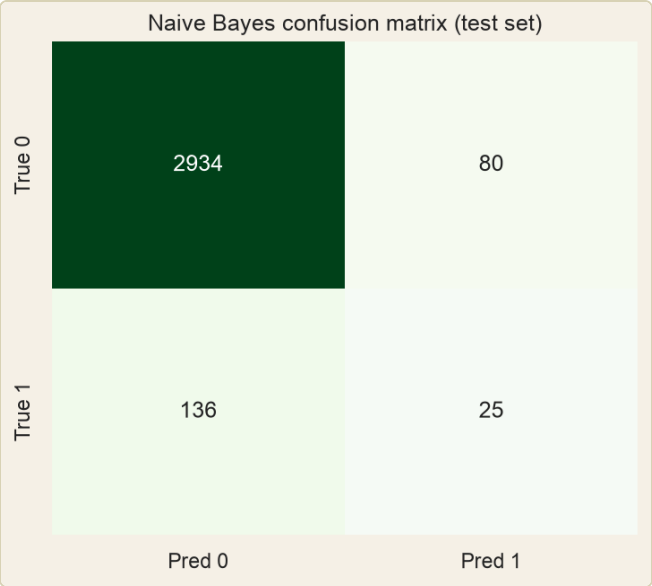

Figure 8. Naive Bayes confusion matrix, test set.

4.3. Why Naive Bayes Underperforms

Naive Bayes' out-of-fold predicted probabilities during cross-validation were severely saturated: a median predicted probability of 0.99998, with 71% of all predictions above 0.9999. This is the concrete, empirical mechanism behind Section 4.1's finding: because Naive Bayes multiplies each feature's individual evidence together as though the features were independent, the two disclosed correlated pairs (customer count with EENS, and historical burning rate with EENS) each get counted twice, pushing predicted probabilities toward the extremes and making the model's confidence far less trustworthy than its raw accuracy figure (93.2%) suggests. Random Forest does not share this failure mode: tree splits do not assume feature independence, so the same correlated pairs that degrade Naive Bayes cause it no equivalent penalty. This is presented as an intentional, disclosed illustration of why checking a model's underlying assumptions matters before trusting its results, not as a hidden weakness of the analysis.

4.4. Comparison Against Published Results on This Dataset

A benchmark comparison against published literature was deliberately deferred until after the first two models above were finalised and tested, to avoid a known external result biasing feature or model choices during development. One prior published study is confirmed to use

this exact dataset and is directly comparable: Alvarez, Lozano, and Bravo Montenegro (*Ingenieria*, 2022) built a support vector machine (SVM) per year, training on a set constructed from every burned transformer plus a subset of healthy ones (2019: 2,417 training rows against 13,452 held out), rather than the natural 5.08% failure rate. That model reached 95.43% accuracy and 56.1% recall in 2019, and 52.4% recall in 2020. The initial logistic regression reached 41.0% recall at a 5.9% false-positive rate, roughly 11-15 percentage points below this verified benchmark. Two disclosed, deliberate methodology differences accounted for that gap, not a modelling error: (1) SVM is a nonlinear model, while the first two models compared were classical linear/probabilistic families by design (Section 3.1); (2) the benchmark study manually undersampled the healthy-transformer class for training, while this analysis used class-weighting alone (Section 3.4) on the natural, unresampled class distribution. Random Forest was added specifically to close the first of these two gaps: as a nonlinear model family, it narrows the shortfall to 9-13 percentage points (42.9% recall versus 52.4-56.1%). One methodology difference remains unaddressed: none of the three models tested here used a resampled (undersampled or oversampled) training set, and this is logged as the concrete, unimplemented next step most likely to close the remaining gap further. A separate, more recent study on the same dataset (Lopez Hernandez et al., 2025), combining synthetic oversampling (SMOTE-Tomek), a loss function tuned for hard-to-classify cases (focal loss), and a Bayesian-optimised neural network, reaches 49% recall at a 1% false-positive rate and 85% at a 10% false-positive rate; given its materially greater modelling sophistication, it is reported here as a further point of reference rather than a like-for-like comparison.

Recommendations

1. Prioritize Random Forest's highest-risk transformers for physical inspection first, since it concentrates the most true failures in the smallest targeted subset of the three models tested.
2. Treat lightning exposure and network length as the two strongest, most actionable risk signals when planning preventive maintenance, per the interpretable logistic regression coefficients.
3. Before relying on Naive Bayes for any operational decision, address its disclosed independence-assumption violations (Section 3.2) or use Random Forest or logistic regression instead.

Limitations and Caveats

The following limitations apply to this analysis:

Location direction unconfirmed: the location predictor is confirmed to be a binary urban/rural indicator, but which numeral (0 or 1) means 'urban' could not be independently confirmed between conflicting published sources; the model uses it as an unordered flag rather than asserting a direction.

Naive Bayes independence violations retained by design: two correlated predictor pairs (customer count with EENS, and historical burning rate with EENS) remain in the Naive Bayes feature set after the two largest multicollinearity pairs were resolved, causing the probability saturation described in Section 4.3. This was a deliberate choice to preserve an identical, directly comparable feature set with logistic regression, not an oversight.

Naive Bayes Gaussian assumption violated for one predictor: the historical burning rate is severely zero-inflated (75-77% exact zeros in both outcome classes) and does not meet Naive Bayes' per-feature normality assumption; it was kept for domain relevance and because dropping it for one model only would break the like-for-like comparison.

Random Forest not hyperparameter-tuned: Random Forest used reasonable default settings (300 trees, a minimum of 5 samples per leaf), not a grid or random search; a tuned Random Forest would likely score higher still. This report demonstrates the directional lift available from a nonlinear model family, not an exhaustively optimised ceiling.

Random Forest uses a different predictor set than logistic regression and Naive Bayes: Random Forest was built on the full ten-predictor candidate set (Section 3.3), while logistic regression and Naive Bayes use a reduced, eight-predictor set to resolve multicollinearity/independence violations Random Forest does not have. This is a deliberate, logged difference, not an oversight, but it means Table 1 and Table 2's Random Forest column is not a strictly apples-to-apples feature-set comparison against the other two models.

No event-level transient or harmonic data: switching surges, lightning strikes at the individual-event level, and harmonic distortion are known real-world drivers of transformer degradation, but this dataset records only zone-level lightning climatology, not per-event transient magnitude, switching-surge counts, or harmonic distortion measurements. A fuller model would benefit from this data if it became available.

Survival analysis was not possible: no installation date, equipment age, or duration field exists anywhere in this dataset, only an annual binary failure flag, which rules out a time-to-event modelling approach regardless of preference.

2020 not yet used for validation: the second year of data was deliberately held out as a future out-of-time robustness check (does risk ranking learned on 2019 still hold a year later) rather than being used in this analysis.

Narrowed but unresolved gap against the verified same-dataset benchmark: a published support vector machine model on this identical dataset reaches 52-56% recall versus Random Forest's 42.9%, a 9-13 point gap (down from 11-15 points for the initial logistic regression). Adding a nonlinear model family closed one of the two originally disclosed methodology differences; the

remaining difference, that the benchmark study manually undersampled its training set and none of the three models here did, is a concrete, unimplemented future-improvement direction, not a modelling error.

Colombian dataset, not a Canadian utility finding: this is a public Colombian utility dataset used as a methodology proxy. No claim is made about any Canadian utility's actual equipment, assets, or reliability.

Exact duplicate rows retained: 78 exact duplicate rows (about 0.5% of the 2019 data) were kept rather than dropped, since no unique transformer identifier exists to confirm whether they represent data-entry errors or genuinely distinct transformers with identical recorded specifications.

AI Usage Disclosure

This report's production and formatting were completed with AI-assisted tools. All methodology, architecture, and analytical review were completed by the author.