Working Paper

What Models Learn from Missing Data: Imputation as Feature Engineering

Rebecca Whitworth, PhD

Independent Researcher

rebeccawhitworth@gmail.com | github.com/r-whitworth

November 2025

Disclaimer. The opinions expressed in this paper are solely those of the author and do not represent the views of any institution, employer, or organization.

Abstract

Standard model validation focuses on the learner but ignores the imputation. We show this creates hidden model risk: two models with identical features and test-set performance can behave completely differently because they learned different replacement rules. Using controlled experiments on mortgage data, we demonstrate that imputation strategy is a first-order modeling choice, not a preprocessing detail.

Using a controlled reconstruction experiment on the 2022 HMDA dataset, we examine how three canonical missingness mechanisms (MCAR, MAR, MNAR) interact with standard imputation methods and learning algorithms. Across learners, the model does not recover the true data-generating process—it traces the surface implied by the imputation. Constant fills collapse variation, linear imputations flatten curvature, and tree-based imputations amplify discontinuities. Apparent stability reflects the fidelity of the backfill, not the integrity of the information.

These findings imply that imputation is a first-order modeling choice, not a preprocessing detail. Validation frameworks should document imputation strategy, re-test sensitivity to alternative reconstructions, and monitor shifts in missingness patterns over time, as each directly alters the model's effective decision surface.[1]

Keywords: Missing data, Imputation, Model behavior, Information geometry, Machine learning, Credit risk, HMDA, Data quality, Model stability, Proxy learning

JEL Codes: C55, C38, C87, G21

All code, figures, and replication notebooks are available at https://github.com/r-whitworth/missingness-imputation.

[1] All code is fully deterministic and reproducible. The author confirms that yes, computers do in fact produce the same output when given the same input and random seed. This revelation was less exciting at 3 a.m. than anticipated. As always, all mistakes are regrettably my own.

Introduction

When validating a credit model, practitioners routinely examine feature importance, test-set performance, and disparate-impact metrics—but rarely audit the *imputation strategy*. Yet two models trained on identical covariates and achieving identical accuracy can exhibit completely different decision surfaces, simply because they filled missing data in different ways. The model that appears stable may, in fact, be learning the replacement rule rather than the underlying relationship.

Missing data are common in applied modeling but are seldom treated as part of the modeling problem itself. Standard practice focuses on how to fill the gaps rather than on what those gaps mean for the learner. In credit modeling and other structured domains, missingness is often systematic—driven by reporting conventions, product design, or applicant behavior—yet most empirical treatments assume it is random or ignorable. The distinction matters: when data are not missing at random, the model learns from the pattern of absence as much as from the observed values.

In mortgage lending specifically, missingness is not merely a data quality issue—it is a market structure issue. Borrowers who omit income documentation are making an economic choice. Lenders who accept or encourage such omissions through "low-doc" or "stated-income" products are making one as well. These patterns of disclosure and non-disclosure correlate with credit risk, and when we impute them away, we erase the economic signal.

The literature on missing-data mechanisms follows the formal taxonomy introduced by [14]: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Subsequent work has emphasized estimation consistency under each condition, focusing on parameter bias or asymptotic efficiency ([12]). Less attention has been paid to how contemporary machine-learning learners—trees, neural networks, or hybrids—behave under these mechanisms, particularly when imputation is performed before model training. Most applied frameworks treat imputation as a preprocessing step external to the learner, despite the fact that the chosen fill strategy reshapes the feature space the model ultimately sees.

Recent studies in explainability and model risk highlight that preprocessing choices alter the geometry of decision boundaries (e.g., [11]; [3]). However, missingness remains a blind spot in these discussions: its effects are often attributed to "noise" or "data quality" rather than to systematic transformation of the input manifold. This paper takes an explicitly model-centric view. Rather than comparing imputers in isolation, it examines how different learners respond to controlled patterns of missingness and reconstruction, using a large, structured dataset (HMDA 2022, [4]) where the true distribution is known before deletion.

We treat imputation not merely as a statistical repair but as a feature-engineering operation—one that redefines the shape of the feature space and, in doing so, alters what the learner can perceive. This view aligns with the perspective of information geometry, which emphasizes that preprocessing steps effectively remap the model's parameter manifold ([1]). In effect, every imputation is a form of feature engineering ([9]), and every pattern of missingness is a form of labeling. When absence correlates with outcomes—such as when missing income coincides with declined applications—the learner internalizes that structure. It does not infer bias; it learns geometry.

Consider a simple example. Suppose income and debt-to-income ratio (DTI) are imputed with their sample means, and borrowers with missing values are disproportionately associated with denials. After imputation, those missing entries now occupy the same "average" region of feature space as a borrower with genuinely average characteristics. Because that region is enriched with denied observations, the learner interprets the geometric center of the distribution as a high-risk zone. When an average borrower subsequently applies, their coordinates fall into that same region, and the model responds accordingly: the ordinary case becomes the mistaken proxy for the omitted one. Similar dynamics have been documented in fairness and bias research, where outcome-correlated missingness (MNAR) systematically distorts classification boundaries ([2, 7]).

This framing motivates the analysis that follows. If preprocessing choices reshape the manifold, then model behavior cannot be evaluated solely by accuracy—it must also be understood as a response to the geometry we impose through missingness and its repair.

Data Handling

The source data are drawn from the public 2022 Home Mortgage Disclosure Act (HMDA) release. To ensure comparability and avoid drawing observations with different underwriting regimes, we restrict attention to conventional, first-lien, site-built, closed-end originations—applications that correspond most directly to the consumer credit decision process and represent the first time the borrower is attempting to access the credit market for this property.[2] Only loans with fully observed income, debt-to-income ratio (DTI), loan amount, age, and region are retained for baseline construction. After filtering, the working dataset contains approximately 5.77 million observations.

Variables were cleaned and standardized following HMDA documentation. Income was reported in thousands and capped at $2 million to remove obvious outliers. DTI was converted from categorical bins to midpoint estimates (e.g., $40\breve{}50\% \rightarrow 45$). Age was converted from categorical ranges to midpoints (e.g., $25\breve{}34 \rightarrow 29.5$).[3] Regions were defined from state codes using four standard Census divisions. Observations with implausible or missing values after cleaning were dropped prior to sampling.

All experiments operate on repeated random subsamples of 100,000 observations drawn without replacement from this clean population. Each Monte Carlo iteration applies the same sampling, deletion, and imputation steps under deterministic seeds to ensure reproducibility. Credit score variables are not

[2] Roughly speaking. That data certainly includes some loan shopping, but the majority of that is swept in removing withdrawn and incomplete applications.

[3] This clearly removes original variation but is functionally limited by what the public lar file provides.

available in the public release and therefore excluded from all analyses.[4][5]

Descriptive Stability

Table 1 compares the full 5.77-million-loan population to the average of the random 100,000-observation draws used in the Monte Carlo experiment.[6] The summary statistics confirm that the subsamples remain representative: means, variances, and quantiles differ by less than one percent across all major covariates. In other words, sampling does not materially distort the baseline distribution of income, loan amount, or DTI. Any subsequent shifts in model metrics can therefore be attributed to missingness and imputation rather than to sampling bias.

Table 1: Summary statistics for the cleaned 2022 HMDA dataset (Panel A) and for the 25 Monte Carlo subsamples of 100,000 loans each (Panel B). The near-identical moments confirm that experimental draws preserve the full-sample structure.

Panel A. Full cleaned sample (5.77 million loans)

Variable	Mean	SD	Min	Median	Max	Missing (%)
Loan amount ($)	396,452	1,942,976	5,000	275,000	1,491,565,000	0.0
Income (000s)	140.51	163.28	-133.0	102.0	2,000.0	5.6
Debt-to-income (%)	36.38	12.33	10.0	37.0	70.0	8.6
Applicant age (years)	46.36	14.11	22.5	49.5	77.5	3.6
Denial indicator (0/1)	0.149	0.356	0.0	0.0	1.0	0.0

Panel B. Monte Carlo subsample stability (25 draws of 100 K each)

Variable	Mean ± SD of means	SD ± SD	Median ± SD
Income (000s)[7]	140.5 (± 0.2)	163.3 (± 0.3)	102.0 (± 0.0)
Debt-to-income (%)	36.43 (± 0.04)	12.26 (± 0.03)	37.6 (± 0.49)
Applicant age (years)	46.27 (± 0.04)	14.12 (± 0.03)	49.5 (± 0.0)
Denial indicator (0/1)	0.143 (± 0.001)	0.350 (± 0.001)	0.0 (± 0.0)

Research Set Up

To understand how different patterns of absence interact with learning algorithms, we constructed missingness artificially within a clean subset of the 2022 HMDA dataset. This allows a controlled comparison of imputation strategies under known mechanisms to a clear baseline, rather than incidental reporting artifacts. We expose three learners — logit, XGB, and NN — to each type of missingness with each imputation strategy. By carefully stepping through data with a known baseline, we can answer the question **how does missingness and imputation affect the model?**

[4] Credit score fields are suppressed in the public HMDA dataset; all experiments are conducted using only variables available in the public file. There is always concern over omitted variables causing bias in the estimates.

[5] All processing and model estimation were conducted in `Python 3.11` using `scikit-learn 1.5`.

[6] Monte Carlo methods are well known for verifying the convergence of small samples to population metrics. [5, 6, 8, 13] Here they provide a reproducible baseline for model performance under controlled distortions.

Constructing Missingness

Missingness in observational data is rarely random. We selected two covariates that differ sharply in how easily their signal can be recovered by a model: **income** and **debt-to-income ratio (DTI)**. Income is highly correlated with other observable variables such as loan amount, property value, and geographic indicators. When income is removed, the learner can still infer an approximate value from those related covariates. The signal is diffuse but recoverable: higher loan amounts and larger property values typically correspond to higher reported incomes. In contrast, DTI is only weakly correlated with other inputs. It reflects a borrower's internal balance sheet—liabilities and cash flow—not something the lender can easily predict from external file data. When DTI is missing, the learner faces a genuine information gap rather than a reconstructable one.

To study how models respond to these two classes of absence, we imposed three canonical forms of missingness: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). These formal categories map closely to real-world reporting behaviors. MCAR corresponds to mechanical loss—records dropped during transmission or fields left blank at random. MAR captures structured but explainable gaps, such as lower-income applicants being less likely to report income or certain lenders omitting DTI for specific product types. MNAR represents the hardest case: data that go missing precisely when their values are problematic, as when borrowers with high DTIs choose not to disclose them or lenders redact unfavorable ratios.

We intentionally designed the subset of HMDA data to be complete for these variables. Missingness was created synthetically to study the downstream implications relative to the learner's baseline with fully known covariates. For each mechanism, we began with a complete dataset and randomly deleted a fixed proportion of values according to the specified rule (uniform for MCAR, covariate-linked for MAR, outcome-linked for MNAR). This approach isolates the effect of missingness itself: the underlying distribution of true values remains known, allowing a direct comparison between what the learner *should* see and what it *does* see after imputation.

All deletions were applied before model training to prevent leakage from downstream transformations, and each variant was later imputed using one of three fill strategies (mean, linear, or tree). This design allows the learner's behavior to be observed under controlled, interpretable distortions that mirror practical data-preparation choices.

Missing Completely at Random (MCAR). Under MCAR, entries were deleted independently of all other variables. For each draw, a fixed proportion p of values in the target column (income or DTI) was replaced with missing indicators using a uniform random mask:

$$M_i \sim \text{Bernoulli}(p), \quad X_i^{\text{obs}} = \begin{cases} X_i & \text{if } M_i = 0, \\ \text{NA} & \text{if } M_i = 1. \end{cases}$$

This provides a neutral benchmark in which missingness has no relation to borrower characteristics or outcomes. Any subsequent degradation in model performance therefore reflects the learner's sensitivity to the imputation strategy itself rather than to structural bias.

Missing at Random (MAR). In the MAR condition, the probability of missingness depends on other observed covariates but not on the unobserved value being deleted. In the implementation, DTI values were dropped with a probability linked to standardized income, such that higher-income observations were more likely to be deleted. This creates structured but explainable missingness—analogous to data-entry or reporting practices that vary systematically across borrower segments yet remain observable to the analyst. Formally,

$$P(M_i = 1) = \text{logit}^{-1}(\alpha + \beta Z_i),$$

where Z_i represents the conditioning variable (here, income). This form of missingness can, in principle, be modeled away, since the mechanism is fully determined by known features.

Missing Not at Random (MNAR). For MNAR, missingness depends on the unobserved value itself. DTI values were selectively deleted among higher-DTI observations according to a logistic weighting function of the standardized DTI score:

$$P(M_i = 1) = \text{logit}^{-1}(\gamma + \delta Z_i),$$

where Z_i denotes the normalized DTI. This design mirrors the practical case in which borrowers with high DTI ratios are less likely to disclose them, or lenders redact such values from applications. The resulting pattern persists in the data but cannot be reconstructed from the observed covariates, producing a hidden dependency that the learner cannot directly observe.[8]

Each mechanism was applied across 25 independent redraws to stabilize variance. The resulting datasets preserve the marginal distributions of non-deleted features while isolating the learner's response to the pattern of missingness itself.

Imputation Strategies

Each missingness condition was paired with one of three imputation strategies that represent common practice in applied modeling: a constant fill (*mean*), a linear model–based fill (*linear*), and a tree–based fill (*tree*). These were selected not for novelty but to illustrate how even standard imputation choices alter the geometry of what the learner perceives as data.

[8] A fourth variant—linking the probability of missingness to the model-implied denial score rather than to the covariate itself—was considered but not implemented. Such an outcome-linked MNAR mechanism would simulate selective disclosure (e.g., lenders omitting DTI for denied loans) and remains open for future work.

Mean imputation. Missing values were replaced with the mean of the observed training data for that variable.[9] This approach collapses the distribution of the affected feature into a single mass point, reducing variance and generating an artificial cluster in feature space. To flexible learners, this cluster appears as an (in)coherent subpopulation rather than as a placeholder. Median and zero–value imputations, which are functionally equivalent to constant fills, behave identically in this respect.

Linear imputation. Here, missing values were estimated from a regression fitted on the observed portion of the data:

$$\hat{X}_i = \alpha + \mathbf{Z}_i^\top \boldsymbol{\beta},$$

where $\mathbf{Z}_i$ represents the non–missing covariates. This approach preserves the global linear relationships among features but smooths over local variation. When the correlation structure is strong (as with income), the recovered values remain close to the true surface; when it is weak (as with DTI), the fitted surface flattens the geometry and attenuates curvature.

Tree–based imputation. Nonlinear imputation was implemented using a gradient boosting regressor trained on the observed data. The model predicts missing values from the remaining covariates, capturing higher–order and local interactions that linear models cannot. This produces fills that are locally adaptive but may amplify existing discontinuities if the underlying data are sparse or segmented. As with the linear case, the effectiveness of this strategy depends on how recoverable the target variable is from the rest of the feature set.

For each imputation method, an additional experiment was run including a binary flag indicating whether the original value had been missing. The flag marginally reduces variance, *but that is about all it does.* The flag does not restore the lost relationship or supply meaningful information about what was missing; it merely labels the absence.

Learners and Evaluation Metrics

We evaluated three learners of increasing flexibility: a logistic regression (**Logit**), a gradient-boosted tree (**Tree**), and a shallow feed-forward neural network (**NN**). Each learner represents a different inductive bias and therefore a different way of interacting with missingness and imputation.

Logistic regression. The logistic model provides a linear baseline. It assumes a single global decision boundary and cannot reconstruct nonlinear relationships. Any degradation in its performance directly

[9] Mean imputation is not materially different from median or zero fills. All three collapse heterogeneous observations into a single value, effectively creating a spike at the center of the distribution. From the learner’s perspective, cases that originally fell on opposite sides of the decision boundary become indistinguishable in this covariate, and the model simply goes ’welp’—treating the cluster as an ambiguous region of low confidence.

reflects loss of separability in the feature space[10]. Because it cannot overfit small artificial clusters, it serves as a reference for how much real information was destroyed versus how much structure was fabricated.

Gradient-boosted tree. Tree-based learners adapt locally. They can exploit discontinuities created by imputation, sometimes recovering apparent accuracy by partitioning around the new mass points.[11] This makes them particularly revealing: if performance is stable despite extensive missingness, it is usually because the model has learned from the pattern of missingness and imputation itself, rather than the underlying data-generating process.

Neural network. The neural network occupies a middle ground. It can approximate nonlinear structure without the hard partitioning of trees, but it remains sensitive to the smoothness of the input manifold. When imputations flatten variation (as with mean fills) or introduce sharp synthetic boundaries, the learner cannot trace local curvature and stability falters. The network therefore provides a consistency check on whether the observed effects are due to data geometry or model bias.

Evaluation. Model behavior was summarized using three complementary statistics: the area under the ROC curve (AUC), coefficient of determination (R^2), and an empirical curvature measure (C) computed from the second derivative of the predicted probabilities. Together, these capture discrimination, explained variance, and the smoothness of the decision surface.

Metrics were averaged across 25 Monte Carlo redraws of 100,000 observations each (with replacement) to stabilize estimates and expose systematic rather than sample-specific effects. Because each run begins from a 100,000-observation draw of the full 5.77-million-loan dataset, within-draw differences in learner performance reflect only the interaction between the learner and the missingness pattern. To account for sampling variability, the entire process was repeated twenty-five times with independent subsamples, and results are reported as draw-averaged estimates.

Together, these configurations define a controlled environment for observing how models respond when information is removed and then synthetically restored. Across learners, imputations, and missingness mechanisms, the experiment isolates two questions: what the learner can genuinely recover, and what it does with the information we backfill in the data manifold. The logistic regression exposes the loss of separability, the tree reveals the creation of synthetic structure, and the neural network traces the boundary between the two.

The next section presents these effects empirically—how the geometry of the model surface shifts as data vanish, and what, if anything, the learner appears to remember.

[10] To the logit, every feature space looks like a hyperplane and every tool looks like a ruler.

[11] To a tree, every landscape looks like a slope waiting to be partitioned. Stability is achieved not by following the terrain, but by partitioning it.

Results

Baseline Performance

Table 2 reports model performance on the complete data (no missingness) using a 100,000-loan random subsample.[12] The hierarchy is stable across all metrics: the logistic model provides a linear baseline, while both the tree and neural network recover additional nonlinearity, reflected in higher AUC, R^2, and curvature.

Table 2: Baseline model performance on complete data (100,000 observations).

Learner	**AUC**	R^2	**Curvature**
Logit	0.678	0.151	0.048
Tree	0.756	0.278	0.060
NN	0.752	0.274	0.061

Even with this limited (publicly available) covariate set, flexible learners achieve nontrivial discrimination ($AUC \approx 0.75$). This indicates that acceptance decisions in the sample (conventional, first-lien, site-built, origination loan applications) are strongly structured, and that much of the predictive signal arises from coarse, publicly observable characteristics rather than detailed applicant information.[13] Subsequent experiments therefore isolate how missingness interacts with that already high baseline separability on the same set of covariates.

The baseline confirms that all three learners converge on a smooth, monotonic decision surface. The tree and neural models show slightly higher curvature and slightly strong recoverability, consistent with local adaptivity rather than noise amplification. This establishes a clear benchmark: subsequent declines in these metrics represent information loss rather than sampling instability.

Predictability of Income and DTI

Before inducing missingness, we evaluated how predictable each target variable is from other observed covariates (loan amount, age, and region). Results are shown in Table 3. Income is moderately reconstructable ($R^2 \approx 0.35$–0.39, AUC ≈ 0.81–0.83), while DTI is effectively idiosyncratic ($R^2 < 0.03$, AUC ≈ 0.55–0.58). This observation becomes key for later results.

Missingness in income has very little impact on the learner, whereas missingness on DTI is more problematic. In the absence of credit score, DTI effectively functions as a *memory* of past credit decisions

[12] The subsample was drawn deterministically using the same random seed as the first iteration of the Monte Carlo experiment (`SEED = 4242`), ensuring consistency between the baseline and subsequent missingness runs.

[13] Credit score variables are unavailable in the public HMDA release; all analyses are therefore conducted without them. In practice, credit score serves as a latent correlate of several included covariates (income, DTI, and region), so the omission affects scale and some wobbliness around credit history, rather than direction of results. The analyses rely solely on variables available in the public release, ensuring that every result could, in principle, be replicated by an external reviewer.

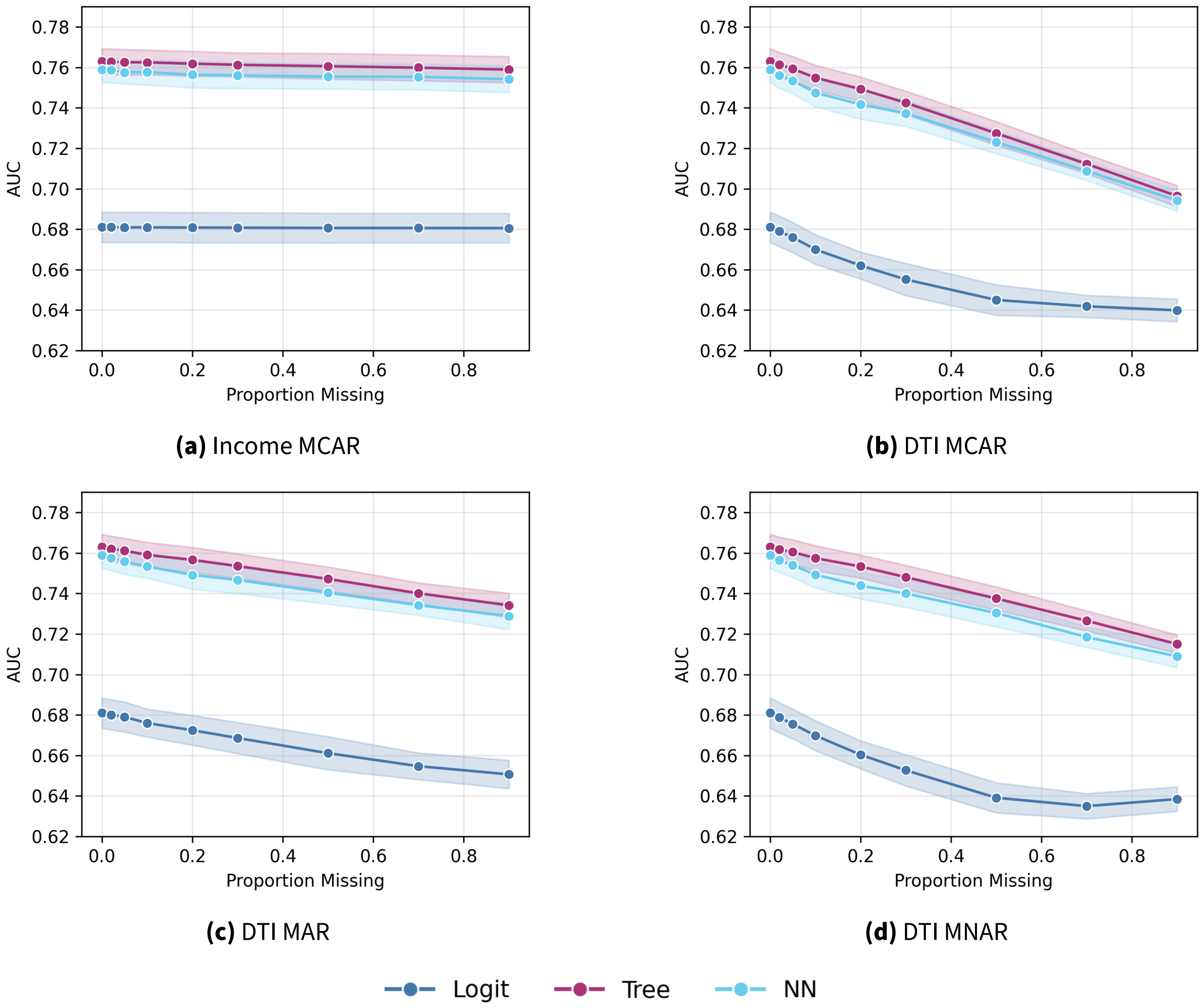
Table 3: Predictability of income and DTI from observable covariates (100,000-observation sample).

Target	Model	R^2	AUC (binary)
Income	Linear	0.346	0.812
Income	Tree	0.388	0.826
DTI	Linear	0.008	0.550
DTI	Tree	0.024	0.581

(cf. [10]); the covariate carries historical information that the learner cannot reconstruct from current observables. In this sense, DTI summarizes the borrower's cumulative credit risk—information that is orthogonal to other reported features. As the number of redundant explanatory variables increases, the impact of missingness in any one of those variables should shrink. These differences set the stage for the experiments: when income is removed, the learner can approximate it from correlated features; when DTI is removed, the model faces a genuine loss of information—no feasible reconstruction exists.

Mean Imputation

Mean imputation remains a fixture in production data pipelines because it is simple, fast, and rarely breaks anything downstream. It preserves record counts, avoids NaN propagation, and produces tidy, model-ready matrices—features prized in operational settings more than statistical fidelity. Conceptually, it assumes that missing borrowers resemble the average of those observed, substituting a constant for information that never existed. This collapses genuine variation (including observations from both sides of the decision boundary) into a single mass point. T he result is a compressed feature distribution: distinct applicants are collapsed into a single artificial value, reducing variance while destroying separability. To the learner, the filled cluster does not look neutral—it looks like a small, (in)coherent subpopulation. Figure 1 shows how this plays out empirically across the three missingness mechanisms (MCAR, MAR, MNAR) for income and DTI.



(a) Income MCAR

(b) DTI MCAR

(c) DTI MAR

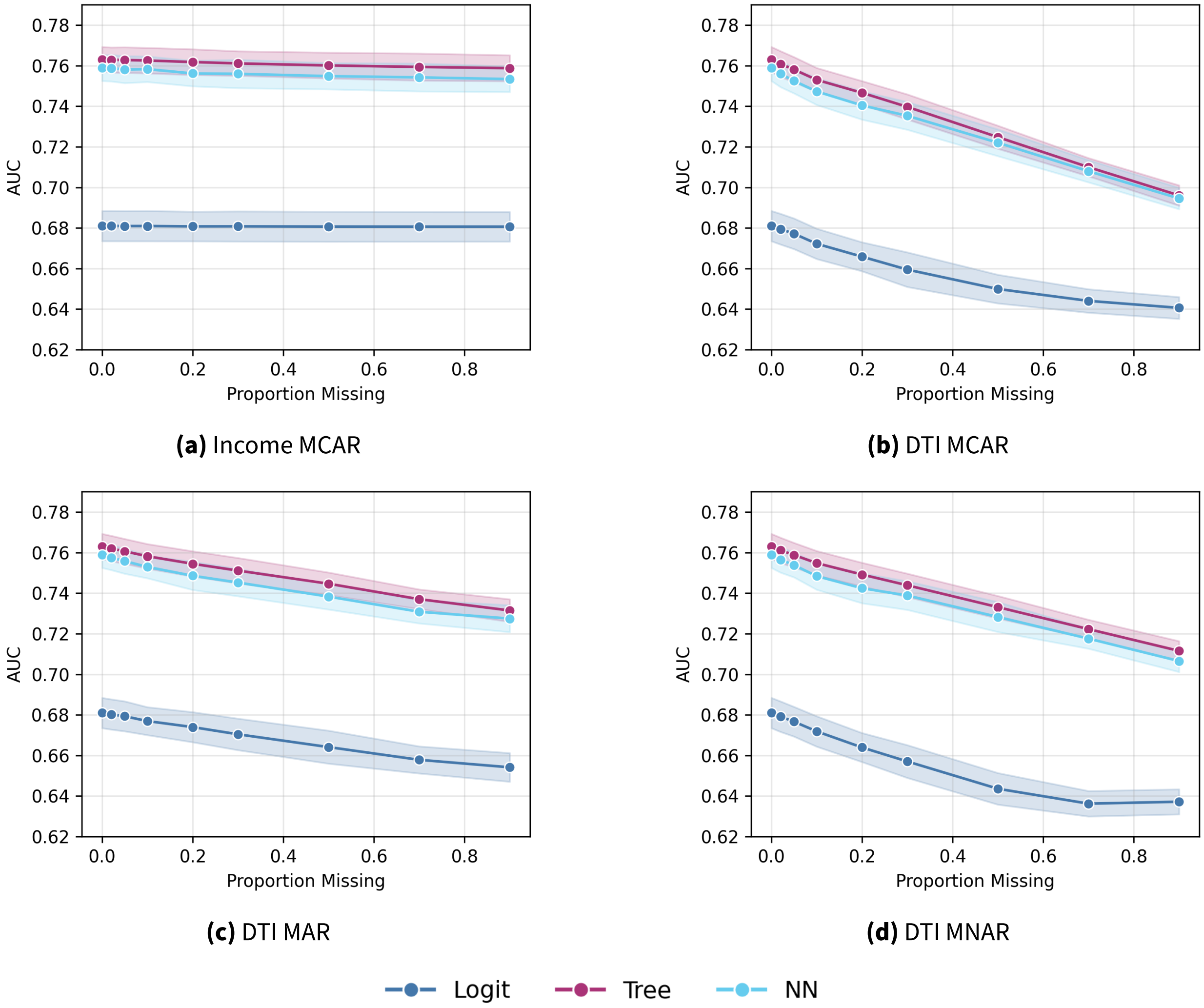
(d) DTI MNAR

Figure 1: Mean-imputation performance under different missingness mechanisms. Lines show AUC averaged across twenty-five Monte Carlo draws (shaded areas denote ± 1 SD). Performance for income (top left) remains nearly constant regardless of deletion rate, reflecting its high correlation with other covariates. For DTI (right and bottom panels), AUC declines linearly with the proportion missing, even under MCAR. All learners converge toward the same flattened trajectory, indicating that constant-value imputations erase variance rather than recover signal.

The models interpret the imputed cluster as a homogenous, low-information region in feature space. Across learners, mean imputation produces the same basic distortion: it lowers overall discrimination while preserving the illusion of stability. Because every missing observation is replaced by the same constant, performance decays smoothly with the proportion missing, even under MNAR. The model cannot recover what was lost, but neither does it visibly fail—it simply learns around the artificial cluster it has been handed. In income, where the missingness is reconstructable, this substitution matters little; in DTI, where the variable carries unique signal, the same operation erases structure that cannot be replaced.

Linear Imputation

Linear regression reconstructs missing information from the global linear relationship among observed covariates, rather than assuming missing borrowers are “average”. It is stable, explainable, and fast. Where strong collinearity exists (that is, the missing signal doesn’t add much new information)—as with income—it can nearly restore the original data-space. Where the relationships are weak or underdefined—as with DTI—it substitutes a flattened projection of the data manifold, preserving mean trends but erasing local curvature and failing to fill information unique to that observation. The resulting surface is smoother than a constant fill, but it remains synthetic: a manufactured gradient that the learner treats as correlation.



(a) Income MCAR

(b) DTI MCAR

(c) DTI MAR

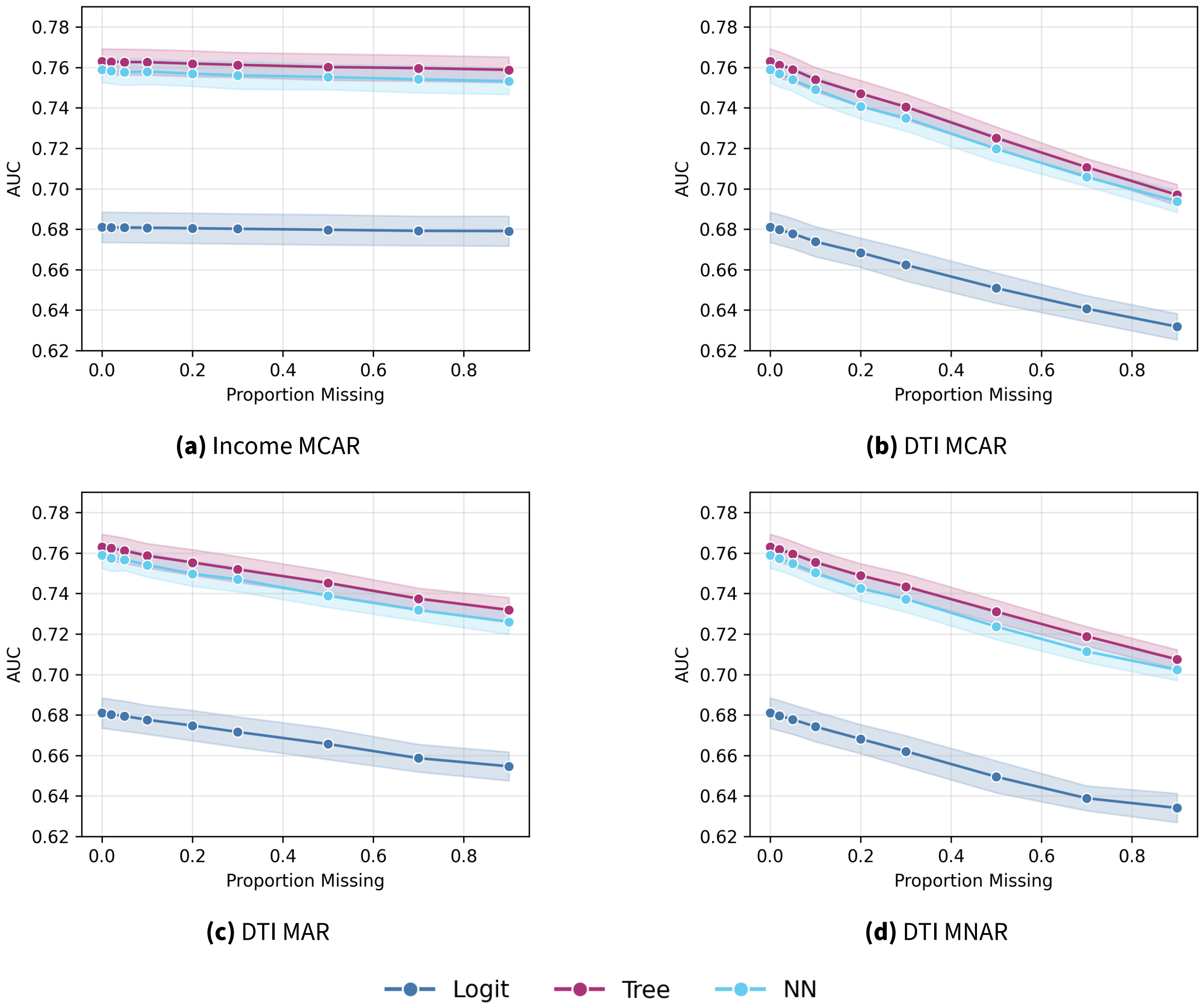
(d) DTI MNAR

Figure 2: Linear (regression-based) imputation across missingness mechanisms. Performance degrades more gradually than under constant fills, especially for income, but DTI remains sensitive. The smoother slopes indicate partial recovery of global relationships, though the flattening in the lower panels shows that weakly correlated variables cannot be reconstructed by linear projection alone.

Linear imputation improves reconstruction only where true linear dependencies exist. In the income experiments, performance remains close to the baseline across all missingness mechanisms, confirming that the learner can recover most of the lost information from correlated features. In DTI, however, the method merely interpolates a correlated covariate—the surface appears stable but carries little truth. AUC and R^2 fall steadily with increasing missingness, yet curvature remains deceptively low because the learner is tracing the imposed plane (the accuracy of the imputed data) rather than the true landscape (the underlying data generating process we are trying to estimate on a broader level).

Tree-Based Imputation

Tree-based imputations extend the same idea—can we recover the missing signal as a combination of the other covariates—nonlinearly. Instead of fitting a single global surface, the regressor partitions the data into local regions and fits small conditional models within each. Where patterns are recoverable, the method adapts quickly and fills missing values with locally consistent estimates. Where the relationships are sparse or weak, it propegates discontinuities already present in the data. In practice, this often yields the highest apparent accuracy among the imputation methods, but some of that gain reflects learning the pattern of missingness itself rather than the missing information. The fills are locally precise but not necessarily meaningful outside their partition.



(a) Income MCAR

(b) DTI MCAR

(c) DTI MAR

(d) DTI MNAR

Figure 3: Tree-based imputation under different missingness mechanisms. Apparent AUC stability masks curvature distortion: the learner fits local discontinuities introduced by the imputer. For DTI under MAR and MNAR, the models maintain accuracy by learning the replacement rule rather than the true relationship, creating synthetic structure.

Tree-based imputations recover most of the structure in income and modestly improve stability in DTI. Performance remains higher than with linear fills at equivalent missingness rates, particularly under MAR and MNAR. However, the improvement is diagnostic rather than substantive: the model is learning the geometry of the imputation, not the population relationship needed for consistent forward looking application of the model. This reinforces the central distinction of the experiment—recovering structure is not the same as recovering information.

Summary

Within learners, the logistic model's performance declines monotonically with increasing missingness but eventually levels off near an AUC of 0.625, indicating that it has reached the information floor available in the remaining covariates. The slower rate of decline under more sophisticated learners reflects partial recovery of missing structure. The tree and neural models show small shifts in curvature and apparent stability as missingness increases, suggesting that flexibility allows adaptation to replacement patterns rather than true resilience to information loss.

Taken together, these results show that model robustness to missingness is mostly illusory and interpretation collides across several different regions.

- **The correlation between missingness and other covariates matters.** Income appears stable because its signal is distributed across correlated covariates; DTI collapses because it is not.
- **The type of imputation matters.** Mean fills suppress variance and collapse disparate observations into a mass point. Linear and tree imputation fills missingness with signal, but the signal is derived from existing covariates and adds no new signal.
- **The learner's own flexibility governs how well it can adapt to the fragmented structure.** More flexible learners are able to adapt to the infill, whereas linear models are unable to maintain separability.

The apparent resilience of complex learners reflects adaptation to the imputation pattern, not recovery of lost information.

Practical Implications for Model Validation

The experiments highlight that imputation strategy shapes the model's effective decision surface: Two models trained on identical covariates but with different imputations are not equivalent in their identification space. For practitioners, the implication is operational rather than theoretical: changes in reconstruction rules change the model itself.

- **Model development:** Choose imputation strategies based on reconstruction quality for specific variables, not convenience. Tree-based imputations may improve apparent accuracy without improving out-of-sample stability.
- **Model validation:** Evaluate model sensitivity to alternative imputations. If performance varies materially across imputers, the model is learning the replacement rule rather than the underlying relationship.
- **Model monitoring:** Track missingness rates and their distribution over time. Shifts in which borrowers are missing data can change model behavior even if parameters remain fixed.

Summary and Implications

Across all experiments, the central finding is simple: missingness interacts with the learner, not just with the data. When the missing variable is highly correlated with others (as with income), imputation methods—however crude—do little harm. When the variable carries unique information (as with DTI), every imputation method imposes a geometry of its own, altering what the learner perceives as structure. The result is not random noise but organized distortion.

Mean fills collapse heterogeneity into a single point, erasing variation. Linear fills restore continuity but flatten curvature, producing smooth but artificial relationships. Tree-based fills adapt locally, yet much of that stability arises from learning the imputation pattern itself rather than the underlying relationship. Across learners, the logistic model shows degradation as information disappears, while tree and neural models remain deceptively stable, tracing the geometry we added rather than the data we removed.

These results highlight a gap between statistical validity and model behavior. Imputation does not merely “fill in” missing values—it reshapes the feature space and, by extension, the decision surface. Understanding this interaction is critical for domains where model interpretability and fairness depend on how information enters the system. Missingness is not just a data-quality issue; it is a modeling choice, and it carries structure of its own.

References

[1] Amari, S.-i. (2016). *Information Geometry and Its Applications*. Springer.

[2] Barocas, S., Hardt, M., and Narayanan, A. (2021). Missing data, imputation, and bias in algorithmic fairness. *Fairness and Machine Learning*.

[3] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

[4] Consumer Financial Protection Bureau (2023). Home mortgage disclosure act (hmda) public data, 2022 lar file. https://ffiec.cfpb.gov/data-publication/. Accessed November 2025.

[5] Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

[6] Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press, Boca Raton, FL.

[7] Hanna, N., Park, Y., and Price, E. (2020). Towards fairness in missing data imputation: An empirical study. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–333.

[8] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

[9] Jadhav, A., Pramod, D., and Ramanathan, K. (2019). A review of missing data handling techniques in machine learning. *Social Network Analysis and Mining*, 9(1):1–23.

[10] Kocherlakota, N. R. (1998). Money is memory. *Journal of Economic Theory*, 81(2):232–251.

[11] Lipton, Z. C. and Saha, J. (2018). The troubling trends in machine learning scholarship. *Queue*, 16(3):45–77.

[12] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

[13] Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341.

[14] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.