MODEL VALIDATION PACK

Home purchase approval model

Documentation of development, performance, fair lending analysis, and alternatives considered, prepared against the validation pillars of SR 11-7 and the technical documentation requirements of EU AI Act Annex IV.

DEMONSTRATION ARTIFACT

This document was generated automatically from a reference implementation operating on public FFIEC HMDA data. It describes a model trained to replicate historical lender approval decisions, not to estimate credit risk, and it is not a validation of any deployed system. See section 1 and section 10.

MODEL VERSION	v002
MODEL HASH (SHA-256)	`5006c9d1dac1af1ba1d874b69f43a0af33caca4cea180771ec7b6fa112124157`
TRAINING POPULATION	CO 2024 · 39,466 applications
OUT-OF-TIME FRAME	CO 2025 · 51,038 applications
DECISION THRESHOLD	0.5
GENERATED	11 August 2026 at 19:25 UTC
ISSUED BY	Meridian Community Lending (DEMONSTRATION)

Contents

1 Model purpose, intended use, and limitations of use

What the model is for, and — stated explicitly — what it is not for.

Intended use

The model produces an approval or decline recommendation for a conventional, first-lien, owner-occupied home purchase application. It returns a probability of approval; applications scoring above 0.5 are recommended for approval. It is intended to operate inside a decision service that records every decision, produces reason codes for every decline, and submits to the fair lending measurement described in section 7.

What this model does not do

It does not estimate credit risk. The target variable is the lender's recorded action — originated or denied — not repayment. The model has never observed a default, a delinquency, or a loss. It learns to reproduce historical approval behaviour, and any bias in that behaviour is part of what it reproduces. Every disparity reported in section 7 should be read as a disparity in approval behaviour, not as evidence about applicants.

It does not see a credit score. Public HMDA excludes it. The most predictive variable in consumer underwriting is absent, so the performance in section 5 is not representative of a production underwriting model, and the apparent importance of the variables that are present is inflated.

It does not see race, ethnicity, sex, or age. Those attributes are held only on the evaluation frame, keyed by row identifier, and a test in the repository asserts that the intersection of the model's feature list and the protected attribute set is empty. Section 7c examines whether the permitted features carry that information regardless.

Population

Applications in CO with action taken of originated or denied, for home purchase, principal residence, first lien, conventional, and not primarily for a business or commercial purpose. Applications outside that perimeter are out of scope and the model must not be applied to them.

2 Data lineage

Source, vintage, every filter applied, and the row count at each stage.

SOURCE	FFIEC HMDA Data Browser — public Loan/Application Register
ENDPOINT	`https://ffiec.cfpb.gov/v2/data-browser-api/view/csv`
GEOGRAPHY	CO
TRAINING VINTAGE	2024
OUT-OF-TIME VINTAGE	2025
EXTRACTED	2026-08-10

Only actions_taken and loan_purposes can be applied server-side: the API caps at two filter categories and has no occupancy, lien, loan type, or business-purpose parameter. The remaining filters are applied locally and their cost is shown stage by stage below.

Row counts at each stage

VINTAGE	STAGE	ROWS	RETAINED	NOTE
2024	raw rows from API	89,988	100.0%	hmda_CO_2024.csv
2024	action_taken in (1,3), purpose=1	89,988	100.0%	applied server-side; verified locally
2024	occupancy_type in (1)	81,343	90.4%	principal residence
2024	lien_status in (1)	76,869	85.4%	first lien
2024	loan_type in (1)	53,587	59.5%	conventional
2024	business_or_commercial_purpose in (2)	53,213	59.1%	not primarily commercial
2024	income and loan amount present	52,478	58.3%	not imputed
2024	modelling frame	52,478	58.3%	CLTV derived for 4,175, still missing for 0
2025	raw rows from API	90,440	100.0%	hmda_CO_2025.csv
2025	action_taken in (1,3), purpose=1	90,440	100.0%	applied server-side; verified locally
2025	occupancy_type in (1)	81,481	90.1%	principal residence
2025	lien_status in (1)	76,305	84.4%	first lien
2025	loan_type in (1)	52,137	57.6%	conventional
2025	business_or_commercial_purpose in (2)	51,717	57.2%	not primarily commercial
2025	income and loan amount present	51,038	56.4%	not imputed
2025	modelling frame	51,038	56.4%	CLTV derived for 3,212, still missing for 0
2024	split: train	39,466	43.9%	2024 development sample
2024	split: test	13,012	14.5%	2024 in-time holdout
2025	split: oot	51,038	56.4%	2025 out-of-time holdout — never seen in training

Exclusions and their rationale

- **Action taken other than originated or denied.** Withdrawn, incomplete, and preapproval outcomes do not represent a credit decision on a complete application and would introduce a label that means something different.
- **Purposes other than home purchase.** Refinancing and home improvement have different underwriting economics; pooling them would produce a model that is well specified for none of them.
- **Non-principal residences, subordinate liens, non-conventional products, and commercial-purpose lending.** Each is underwritten to different standards and different regulatory perimeters.
- **Applications with no reported income or loan amount.** These cannot be scored on the two most important available factors and are dropped rather than imputed.

Split design

hash-partitioned holdout within the training year; the following year is held out entirely as an out-of-time frame.

The public Loan/Application Register carries no date field beyond the activity year, so an intra-year temporal split is not available. Using the following vintage as the holdout is the only construction that yields a genuine out-of-time frame, and it doubles as the drift comparison in section 6. The cost is that the most recent vintage is not used for training, which is a deliberate trade in favour of stability evidence.

SPLIT	ROWS	VINTAGE	APPROVAL RATE	ROLE
train	39,466	2024	87.9%	2024 development sample
test	13,012	2024	87.7%	2024 in-time holdout
oot	51,038	2025	90.0%	2025 out-of-time holdout — never seen in training

3 Features, transformations, and constraints

Each input, how it is derived, and the written rationale for its monotonicity constraint.

The model uses 11 features. Each carries a monotonicity constraint of +1, −1, or 0, and each constraint carries a written rationale reproduced below. A constraint of 0 is a documented decision that no direction can be defended, not an omission. Constraints are verified empirically after training: each constrained feature is swept across its observed range on sampled applications with all other inputs held fixed, and the score is asserted never to move against its constraint.

annual income `log_income` +1 NON-DECREASING

Source field	`income`
Transformation	log1p(income_usd), income capped at the training 99th percentile

Higher income increases capacity to repay. A model in which additional income could reduce the probability of approval would be indefensible to a reviewer and could not be explained to an applicant, so the relationship is constrained to be non-decreasing.

loan amount `log_loan_amount` -1 NON-INCREASING

Source field	`loan_amount`
Transformation	log1p(loan_amount_usd); HMDA reports this midpoint-rounded to $10,000

Holding income and property value fixed, a larger requested loan represents greater exposure and a larger required payment. The relationship is constrained to be non-increasing so that requesting more money can never, by itself, improve the applicant's position.

combined loan-to-value ratio `cltv` -1 NON-INCREASING

Source field	`combined_loan_to_value_ratio`
Transformation	numeric, clipped to [0, 200]

A higher loan-to-value ratio means less borrower equity and a thinner cushion against loss on default. This is one of the most established directional relationships in mortgage credit, so it is constrained to be non-increasing.

debt-to-income ratio `dti_band_ord` -1 NON-INCREASING

Source field	`debt_to_income_ratio`
Transformation	ordinal encoding of the HMDA DTI band string (see schema.DTI_BAND_ORDER)

A higher ratio of monthly debt obligations to income leaves less residual income to absorb payment shock. HMDA reports this as a coarse band rather than an exact figure, so the feature is ordinal; the constraint is non-increasing across bands.

loan term `loan_term` UNCONSTRAINED

Source field	`loan_term`
Transformation	numeric months, clipped to [60, 480]

Loan term has no unambiguous directional relationship to credit quality. A longer term lowers the monthly payment (favourable to capacity) while extending exposure (unfavourable). Because a defensible direction cannot be asserted, the feature is left unconstrained and this is disclosed rather than a direction being invented.

interest-only payment feature `interest_only` -1 NON-INCREASING

Source field	`interest_only_payment`
Transformation	binary; HMDA 1111 (Exempt) treated as missing

Interest-only structures defer principal amortization and are associated with elevated risk. The constraint is non-increasing so the presence of the feature can never improve the score.

balloon payment feature `balloon` -1 NON-INCREASING

Source field	`balloon_payment`
Transformation	binary; HMDA 1111 (Exempt) treated as missing

A balloon payment concentrates repayment risk at maturity and creates refinance dependence. Constrained non-increasing on the same reasoning as interest-only.

negative amortization feature `neg_amortization` -1 NON-INCREASING

Source field	`negative_amortization`
Transformation	binary; HMDA 1111 (Exempt) treated as missing

Negative amortization allows the principal balance to grow, eroding equity over time. Constrained non-increasing.

tract-to-MSA income percentage `tract_to_msa_income_pct` UNCONSTRAINED

Source field	`tract_to_msa_income_percentage`
Transformation	numeric percent, clipped to [0, 300]

Tract relative income is included as an area economic control. No direction is asserted: while area income correlates with local economic conditions, imposing a direction on a geographic variable would embed an assumption about neighbourhood quality that the project is not willing to defend. Its correlation with protected classes is measured explicitly in the proxy scan (section 7).

number of units `units_ord` UNCONSTRAINED

Source field	`total_units`
Transformation	ordinal encoding of the HMDA units band string

Unit count distinguishes single-family from small multifamily collateral. The credit implication is mixed — additional units imply rental income as well as additional exposure — so no direction is imposed.

preapproval request `preapproval` UNCONSTRAINED

Source field	`preapproval`
Transformation	binary (1 = preapproval was requested)

Preapproval indicates a stage of the application process rather than a credit characteristic. It is retained as a process control with no asserted direction.

Excluded by design

tract minority population percentage EXCLUDED

A textbook geographic proxy for race. Excluding it is a modelling choice, not a data limitation; the less-discriminatory-alternative search quantifies what including it would cost in measured disparity.

Protected attributes

derived_race, derived_ethnicity, derived_sex, applicant_age are held out of the model entirely. They exist only on the evaluation frame, keyed by row identifier, and are used solely for the analysis in section 7.

Missing value treatment

LightGBM does not enforce monotone constraints for rows routed through a missing-bin split, and the failure is not confined to the feature that is missing. Eliminating missing values at training is what makes the stated constraints real, and is a precondition for the counterfactual binary search.

FEATURE	SHARE IMPUTED	VALUE USED
dti_band_ord	0.765%	8
loan_term	0.119%	360
interest_only	0.005%	0
balloon	0.005%	0
neg_amortization	0.005%	0

4 Development methodology

Algorithm, hyperparameters, and the measures taken to make training reproducible.

Algorithm

Gradient-boosted decision trees (LightGBM), chosen for its support of per-feature monotonicity constraints. Constraints are the reason for the choice: they make the model's directional behaviour a property of the artifact rather than a claim about it, and they are what makes the counterfactual search in section 9 sound.

Hyperparameters

PARAMETER	VALUE
`objective`	binary
`learning_rate`	0.05
`num_leaves`	31
`max_depth`	-1
`min_child_samples`	100
`reg_lambda`	1.0
`reg_alpha`	0.0
`feature_fraction`	1.0
`bagging_fraction`	1.0
`monotone_constraints_method`	advanced
`deterministic`	True
`force_row_wise`	True
`verbose`	-1
`seed`	20260810
`bagging_seed`	20260810
`feature_fraction_seed`	20260810
`data_random_seed`	20260810
`num_threads`	4
`best_iteration`	157

Reproducibility

Two training runs on the same data and seed produce the same model file. This is not automatic: LightGBM's histogram construction order depends on how work is divided across threads, so `deterministic`, `force_row_wise`, and a fixed thread count are all required. Without reproducibility a decision recorded against “v002” could not be replayed later, because the version string would not identify a specific set of bytes.

MODEL ARTIFACT HASH	5006c9d1dac1af1ba1d874b69f43a0af33caca4cea180771ec7b6fa112124157
TRAINING ROW COUNT	39,466
TRAINING ROW IDENTIFIER HASH	b756113f7599fc4356b933c2c9fd4c0b19d58af83855d653137eaeb8664e0bfb
EXPLANATION BACKGROUND HASH	6479949161a8d160…

The row identifier hash fingerprints the exact training sample without storing it, so a reviewer can establish whether two models saw the same data.

5 Performance

Discrimination and calibration overall and by segment.

AUC · IN-TIME HOLDOUT	AUC · OUT-OF-TIME	KS · OUT-OF-TIME	BRIER · OUT-OF-TIME
0.9019	**0.8764**	**0.657**	**0.0446**
13,012 applications	51,038 applications	maximum separation	lower is better

SPLIT	N	AUC	KS	BRIER	PREDICTED APPROVALS	ACTUAL APPROVALS	ACCURACY
train	39,466	0.9120	0.7132	0.0437	90.5%	87.9%	94.7%
test	13,012	0.9019	0.7184	0.0441	90.2%	87.7%	94.8%
oot	51,038	0.8764	0.6572	0.0446	92.3%	90.0%	94.6%

> READING THESE FIGURES
>
> Discrimination is measured against the lender's historical decision, so a high AUC means the model reproduces past approval behaviour closely. It is not evidence that the decisions being reproduced were correct. Absent a credit score, these figures are also not comparable to a production underwriting model.

Loan size band — out-of-time frame

SEGMENT	N	AUC	KS	PREDICTED APPROVALS	ACTUAL APPROVALS
$250k-$500k	21,662	0.7797	0.4612	96.9%	94.0%
$500k-$750k	13,726	0.7997	0.5313	97.3%	95.3%
< $250k	8,808	0.9227	0.7307	69.6%	69.1%
$750k+	6,842	0.7952	0.4480	96.6%	93.6%

Applicant income band — out-of-time frame

SEGMENT	N	AUC	KS	PREDICTED APPROVALS	ACTUAL APPROVALS
$75k-$150k	19,761	0.8442	0.6036	93.8%	91.3%
$150k-$250k	13,614	0.7662	0.4511	98.2%	95.7%
$250k+	9,583	0.6965	0.3019	98.7%	95.6%
< $75k	8,080	0.9434	0.7933	71.1%	70.7%

County (largest twelve) — out-of-time frame

SEGMENT	N	AUC	KS	PREDICTED APPROVALS	ACTUAL APPROVALS
08059	5,760	0.8062	0.4891	97.2%	94.9%
08031	5,692	0.7585	0.4132	97.5%	94.1%
08035	5,071	0.8162	0.5283	97.7%	95.2%
08041	5,058	0.8728	0.6640	92.2%	90.0%
08005	4,971	0.8307	0.5739	94.9%	92.7%
08001	4,767	0.9313	0.7892	83.3%	82.3%
08123	3,787	0.9418	0.8191	85.3%	84.5%
08069	3,657	0.8865	0.7224	92.4%	90.5%
08013	2,700	0.8126	0.5563	96.0%	93.1%
08077	1,480	0.8835	0.7078	91.1%	90.2%
08101	830	0.8612	0.6428	89.5%	85.7%
08014	761	0.8811	0.7000	92.9%	92.2%

6 Stability and out-of-time validation

How performance holds up on a vintage the model never saw, and how the inputs have shifted.

Performance decay on unseen vintage

The out-of-time frame is the following calendar year and was never used in training or model selection. AUC moves from 0.9019 on the in-time holdout to 0.8764 out of time, a change of -0.0256.

Calibration — out-of-time frame

Applications are ordered by predicted probability and split into ten equal groups. A well calibrated model produces a realised approval rate close to the mean prediction in every group.

DECILE	N	MEAN PREDICTED	OBSERVED	DIFFERENCE
1	5,104	0.2813	0.3111	+0.0298
2	5,104	0.9094	0.9236	+0.0142
3	5,104	0.9490	0.9526	+0.0036
4	5,104	0.9595	0.9632	+0.0037
5	5,104	0.9658	0.9671	+0.0013
6	5,104	0.9692	0.9722	+0.0029
7	5,104	0.9721	0.9722	+0.0000
8	5,104	0.9757	0.9747	-0.0010
9	5,103	0.9794	0.9788	-0.0006
10	5,103	0.9913	0.9857	-0.0056

Input and score drift, 2024 to 2025

Population Stability Index computed with bin edges frozen from the training vintage. Below 0.1 is read as stable, 0.1 to 0.25 as moderate, and 0.25 and above as significant.

FEATURE	PSI	BAND	KS
tract_to_msa_income_pct	0.0048	stable	0.0301
log_loan_amount	0.0039	stable	0.0335
dti_band_ord	0.0035	stable	0.0231
log_income	0.0032	stable	0.0199
cltv	0.0014	stable	0.0135
loan_term	0.0011	stable	0.0094
interest_only	0.0000	stable	—
balloon	0.0000	stable	—
neg_amortization	0.0000	stable	—
units_ord	0.0000	stable	0.0010
preapproval	0.0000	stable	—
Predicted probability	0.0048	stable	0.0188

> WHY NO P-VALUES APPEAR HERE
>
> Alerts fire on effect size, never on p-values. With tens of thousands of observations on each side, every two-sample test is statistically significant regardless of whether the change matters, so a p-value-driven alert fires constantly and is then ignored.

7 Fair lending analysis

Raw disparity, disparity surviving controls, and a scan for proxy variables.

All three analyses are computed on the out-of-time frame (51,038 applications from 2025), which the model has never seen. Protected attributes are held only on this frame and never enter a feature vector.

7a — Raw approval rate disparity

> WHAT THE FOUR-FIFTHS RULE IS AND IS NOT
>
> The four-fifths rule is a screening heuristic drawn from the EEOC Uniform Guidelines on employee selection. It is not the legal standard for credit discrimination and a ratio above 0.80 does not establish compliance. It is reported as a first screen only; the controlled analysis is the substantive one.

Race and ethnicity

GROUP	N	APPROVAL RATE	95% INTERVAL	RATIO TO HIGHEST	SCREEN
NH_White REFERENCE	31,044	96.0%	95.7–96.2%	1.000	above 0.80
Black	664	89.9%	87.4–92.0%	0.937	above 0.80
Hispanic	6,350	73.9%	72.8–74.9%	0.770	**below 0.80**
Asian	2,106	94.8%	93.7–95.6%	0.988	above 0.80
AIAN	160	82.5%	75.9–87.6%	0.860	above 0.80
NHPI	34	*insufficient sample — fewer than 100 applicants; no rate is reported*			
TwoOrMoreMinority	33	*insufficient sample — fewer than 100 applicants; no rate is reported*			

Sex

GROUP	N	APPROVAL RATE	95% INTERVAL	RATIO TO HIGHEST	SCREEN
Male REFERENCE	14,474	91.4%	90.9–91.8%	1.000	above 0.80
Female	11,031	88.2%	87.6–88.8%	0.965	above 0.80

Age

GROUP	N	APPROVAL RATE	95% INTERVAL	RATIO TO HIGHEST	SCREEN
35-64 REFERENCE	27,923	92.7%	92.4–93.0%	1.000	above 0.80
62_plus	4,491	92.7%	91.9–93.5%	1.000	above 0.80
under_25	2,255	80.4%	78.8–82.0%	0.867	above 0.80

7b — Disparity surviving controls

A logistic regression of the model's decision on protected class indicators plus the model's own credit factors. This is the analysis that carries weight: a raw rate gap can be explained by differences in the applications, whereas a coefficient that remains significant after the lender's own stated credit factors are included cannot be.

The controls are exactly the model's feature set, so a surviving disparity is disparity the model produces *beyond what its own stated credit factors explain*. Standard errors are clustered by county. The average marginal effect is reported in percentage points of approval probability, which is the quantity a credit committee can act on; the log-odds coefficient is reported alongside it.

Race and ethnicity — reference group NH_White

GROUP	N	COEFFICIENT	STD. ERROR	P	95% INTERVAL	MARGINAL EFFECT
NHPI	34	*insufficient sample*				
TwoOrMoreMinority	33	*insufficient sample*				
Black	660	-0.268	0.220	2.23e-01	-0.70 to +0.16	-0.87 pp
Hispanic	6,191	-0.737	0.076	3.80e-22	-0.89 to -0.59	-2.69 pp
Asian	2,094	-0.487	0.170	4.17e-03	-0.82 to -0.15	-1.67 pp
AIAN	160	-0.667	0.308	3.03e-02	-1.27 to -0.06	-2.39 pp

Fitted on 39,933 applications with 11 controls; McFadden pseudo R^2 0.5543.

Sex — reference group Male

GROUP	N	COEFFICIENT	STD. ERROR	P	95% INTERVAL	MARGINAL EFFECT
Female	10,918	+0.038	0.066	5.63e-01	-0.09 to +0.17	+0.15 pp

Fitted on 25,246 applications with 11 controls; McFadden pseudo R^2 0.5666.

Age — reference group 35-64

GROUP	N	COEFFICIENT	STD. ERROR	P	95% INTERVAL	MARGINAL EFFECT
62_plus	4,438	+0.339	0.114	2.90e-03	+0.12 to +0.56	+1.18 pp
under_25	2,214	+0.039	0.136	7.74e-01	-0.23 to +0.31	+0.15 pp

Fitted on 34,328 applications with 11 controls; McFadden pseudo R^2 0.5250.

The model's decisions compared with the decisions it was trained on

Fitted on the historical lender decision rather than the model's. Comparing the two separates disparity the model introduces from disparity already present in the behaviour it was trained to replicate.

GROUP	MODEL DECISION	HISTORICAL DECISION	DIFFERENCE
Black	-0.87 pp	-2.78 pp	+1.91 pp
Hispanic	-2.69 pp	-4.55 pp	+1.85 pp
Asian	-1.67 pp	-2.57 pp	+0.90 pp
AIAN	-2.39 pp	+0.43 pp	-2.83 pp

> ON CONTROLLING FOR GEOGRAPHY
>
> Two specifications are reported because the choice is contested. County fixed effects control for genuine local economic variation, and they also absorb geographic discrimination, which is itself a fair lending concern. Neither is presented as the single correct answer.

Applications excluded from the analysis

ATTRIBUTE	ASSIGNED	EXCLUDED	SHARE
race	40,391	10,647	20.9%
sex	25,505	25,533	50.0%
age	34,669	16,369	32.1%

Rows whose reported category does not identify a single protected class (Joint, Free Form Text Only, Not Available) are excluded rather than pooled, because a coefficient on such a category cannot be interpreted.

7c — Proxy scan

Normalised mutual information between each feature, quantile-binned into 20 bins, and each protected attribute. A seeded random column is scanned alongside as a negative control to establish the noise floor.

RANK	FEATURE	IN MODEL	NORMALISED MI	MULTIPLE OF FLOOR	STRONGEST AGAINST	FLAG
1	tract_minority_pct	no	0.0379	44.6×	race	flagged
2	cltv	yes	0.0279	32.8×	race	—
3	loan_term	yes	0.0247	29.1×	race	—
4	log_loan_amount	yes	0.0230	27.1×	race	—
5	tract_to_msa_income_pct	yes	0.0215	25.3×	race	—
6	log_income	yes	0.0187	22.0×	race	—
7	dti_band_ord	yes	0.0160	18.8×	race	—
8	interest_only	yes	0.0134	15.8×	age	—
9	balloon	yes	0.0047	5.5×	age	—
10	units_ord	yes	0.0021	2.5×	age	—
11	preapproval	yes	0.0021	2.5×	race	—
12	neg_amortization	yes	0.0007	0.9×	race	—
—	random negative control	no	0.0008	1.0×	—	*noise floor*

> CALIBRATION OF THE FLAG THRESHOLD
>
> The flag threshold (0.03) is a calibrated heuristic, not a natural constant: NMI magnitudes depend on sample size and binning. It is set so that a known geographic proxy is flagged while ordinary credit features are not. The negative control scores 0.0008, which is the floor this scan can resolve; each feature's multiple of that floor is reported so the calibration can be judged directly rather than taken on trust.

Joint predictability of group membership

The model's feature set predicts an applicant's race and ethnicity group with macro-averaged AUC 0.735. An AUC of 0.5 would mean the features carry no information about group membership. This captures joint proxying — combinations of features that together identify a group even when no single feature does — which the pairwise scan cannot see.

WHAT A FLAG DOES AND DOES NOT ESTABLISH

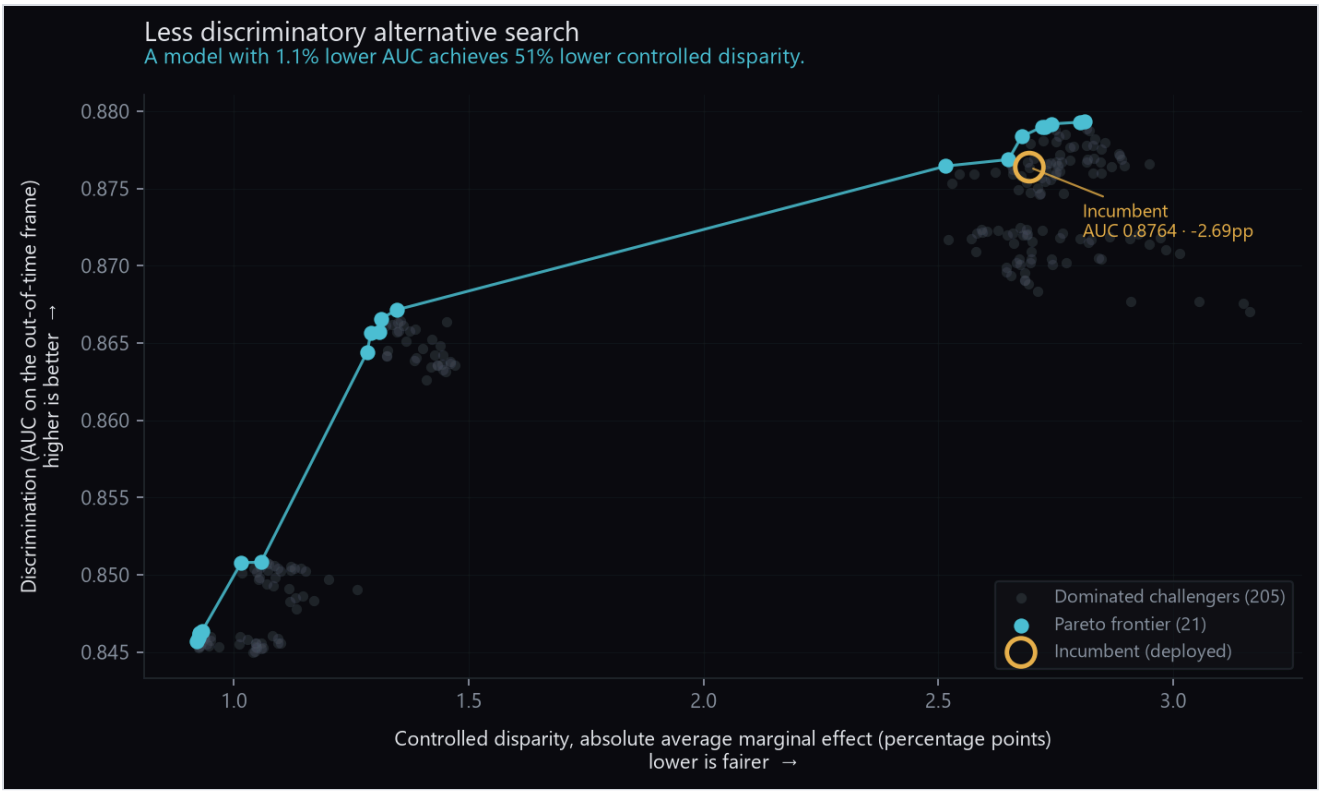
A high value indicates the feature carries information about group membership. It does not by itself establish that the model uses the feature as a proxy — that is what the controlled analysis and the less discriminatory alternative search address.

8 Less discriminatory alternative analysis

A search for a comparably effective model with lower measured disparity.

Under disparate impact doctrine a business-justified model remains challengeable if a comparably effective, less discriminatory alternative exists. This search covers a bounded hypothesis space; the absence of a better alternative within it is not proof that none exists.

> **A model with 1.1% lower AUC achieves 51% lower controlled disparity.**



Each point is a trained challenger evaluated on the out-of-time frame. The horizontal axis is the absolute controlled disparity for Hispanic applicants; the vertical axis is AUC. Points on the frontier are those no other challenger beats on both axes at once. The incumbent is marked.

What was searched

Full factorial over feature subset x fairness weight, plus a seeded random sample of the full hyperparameter space. The full cross product would be 1,440 challengers; this covers the axes that bear on the fairness question exhaustively and samples the rest.

CHALLENGERS EVALUATED	226
FEATURE SUBSETS	baseline, drop_cltv, drop_loan_term, drop_cltv_and_loan_term, drop_geography, minimal, add_known_proxy
LEAVES	7, 15, 31, 63
MINIMUM LEAF SIZE	20, 100, 500
L2 REGULARISATION	0.0, 1.0, 10.0
FAIRNESS WEIGHT	0.0, 0.25, 0.5, 1.0, 2.0
SEARCH SEED	20260810

Objective

Average marginal effect on P(approve) in percentage points, from a logistic regression of the model's decision on protected class plus the incumbent's credit factors, without geographic fixed effects. The control set is held fixed across every challenger so that the axis is a single quantity; controlling each challenger for its own features would reward a model whose proxy absorbs group membership among the controls.

GROUP SEARCHED AGAINST	Hispanic
INCUMBENT DISPARITY	-2.69 pp
STATISTICALLY SIGNIFICANT	yes, p = 3.80e-22

The frontier

MODEL	FEATURE SET	FEATURES	AUC	DISPARITY	APPROVAL RATE	FAIRNESS WEIGHT
Incumbent	baseline	11	0.8764	-2.69 pp	92.3%	—
c071	add_known_proxy	12	0.8794	-2.81 pp	92.3%	0.25
c047	add_known_proxy	12	0.8793	-2.80 pp	92.2%	0.0
c112	add_known_proxy	12	0.8792	-2.74 pp	92.3%	1.0
c033	add_known_proxy	12	0.8790	-2.73 pp	92.4%	1.0
c074	add_known_proxy	12	0.8790	-2.72 pp	92.3%	0.5
c034	add_known_proxy	12	0.8784	-2.68 pp	92.4%	2.0
c139	baseline	11	0.8769	-2.65 pp	92.3%	0.0
c143	baseline	11	0.8764	-2.52 pp	92.3%	1.0
c123	drop_loan_term	10	0.8671	-1.35 pp	93.6%	0.0
c010	drop_loan_term	10	0.8666	-1.31 pp	93.7%	0.0
c063	drop_loan_term	10	0.8657	-1.31 pp	93.6%	0.0
c099	drop_loan_term	10	0.8656	-1.29 pp	93.6%	0.0
c045	drop_loan_term	10	0.8644	-1.28 pp	93.5%	2.0
c016	drop_cltv_and_loan_term	9	0.8508	-1.06 pp	93.8%	0.25
c213	drop_cltv_and_loan_term	9	0.8508	-1.02 pp	93.9%	1.0
c105	minimal	4	0.8464	-0.93 pp	94.0%	0.25
c141	minimal	4	0.8463	-0.93 pp	94.0%	0.0
c026	minimal	4	0.8462	-0.93 pp	94.0%	0.25
c027	minimal	4	0.8462	-0.93 pp	94.0%	0.5
c060	minimal	4	0.8460	-0.92 pp	94.0%	2.0
c199	minimal	4	0.8457	-0.92 pp	94.0%	1.0

Interpretation

The frontier describes the trade available, not the choice to be made. Adopting a challenger means accepting a different distribution of approvals, and the applicants who change outcome are not the same people who changed outcome under the incumbent. That comparison — how many applications flip, in which direction, and for whom — is what a model risk committee needs before it can act, and it is produced separately by the adoption analysis.

Note that including the known geographic proxy as a challenger feature set was evaluated deliberately, in order to price what using it would cost in measured disparity rather than to assert without evidence that it is unacceptable.

9 Explainability and adverse action notices

How reasons for a decline are derived, and a worked example.

Attribution method

Reason codes are derived from Shapley additive explanations computed exactly for tree ensembles. The choice that matters is the reference point.

Explaining an application against the population mean answers “why is this applicant unusual”. Regulation B asks a different question: what would have had to be different for approval. That is a question about the decision boundary, so the reference has to be the boundary. The background dataset is therefore a sample of real applications sitting at the decision threshold, selected so that their mean score equals the cutoff. Contributions then sum to the applicant's distance from the decision boundary rather than to their distance from the average applicant.

Attribution is performed in log-odds space, where the model is additive, and converted to probability only for display. The probability threshold 0.5 corresponds exactly to a log-odds threshold because the logistic function is strictly monotone.

Reason code derivation

The four factors contributing most strongly against approval are selected and mapped to plain language. A raw feature name is never disclosed: every code has applicant-facing text, and the system raises rather than emitting a code without one. Approvals produce no reason codes, since the disclosure duty attaches to adverse action.

Counterfactual derivation

For the single strongest contributor, the system searches for the value of that factor at which the decision would flip, holding everything else constant. The search is a binary search, which is valid only because the model is monotone in that factor — the constraints in section 3 are what make it sound. The resulting value is re-scored and verified before it is disclosed, and it is never permitted to leave the range observed in training. Where no achievable value would change the outcome, the system says so explicitly rather than quoting a number that would mislead.

Worked example

A declined application drawn from the out-of-time frame, scored by model v002.

DECISION	DECLINED
PROBABILITY OF APPROVAL	0.2684
DECISION THRESHOLD	0.5
LEDGER IDENTIFIER	`bc962419-d7ae-4f93-b0af-9051a3cdcb42`

Principal reasons disclosed

#	CODE	APPLICANT-FACING STATEMENT	CONTRIBUTION
1	DTI_HIGH	Your total monthly debt payments are high relative to your monthly income	-1.280
2	LTV_HIGH	The loan amount is high relative to the value of the property, leaving limited equity	-0.452
3	AREA_INCOME_LOW	Economic indicators for the area in which the property is located	-0.137
4	INCOME_LOW	Your income is low relative to the amount you requested	-0.129

Counterfactual disclosed

A debt-to-income ratio at or below 49% would have produced an approval, holding all other factors constant.

Method: ordinal band search. Verified by re-scoring the modified application, which produces a probability of 0.9475 and a decision of APPROVED.

Attribution diagnostic

Monotonicity constrains the model as a function; it does not constrain individual attributions, because interactions can produce a contribution pointing against a feature's overall direction. The table reports how often each constrained feature's contribution agrees with its constraint, over applications where the feature differs materially from the reference. This is reported rather than asserted: a diagnostic that can fail is more informative than a threshold tuned until it passes.

FEATURE	AGREEMENT	APPLICATIONS SCORED	STATUS
log_income	100.0%	752	*ok*
log_loan_amount	62.8%	764	*ok*
cltv	100.0%	750	*ok*
dti_band_ord	100.0%	753	*ok*
interest_only	100.0%	1,460	*ok*
balloon	—	0	*not exercised — feature is constant across this sample*
neg_amortization	—	0	*not exercised — feature is constant across this sample*

10 Known limitations

What this model and this document do not establish.

Reproduced verbatim from the repository's limitations document, so that what appears here is what the project maintains rather than a summary written for this document.

What this model is, and what it is not

The model is trained on `action_taken`, the outcome the lender recorded: originated or denied. It therefore learns to **replicate historical lender approval behaviour**, including whatever bias that behaviour contained. It is **not a default-risk model**. It has never seen a repayment outcome and cannot be used to estimate one.

This distinction matters for how every number in this document should be read. A disparity measured here is a disparity in *approval behaviour*, not evidence about creditworthiness. Improving a fairness metric means the model reproduces the historical pattern less faithfully, which is a defensible goal for a lender but is not the same as being more accurate.

Public HMDA has no credit score

The single most predictive variable in consumer credit underwriting is absent from the public Loan/Application Register. Performance figures here are therefore not representative of what a production underwriting model would achieve, and the apparent importance of the variables that *are* present is inflated, because they are partly standing in for the one that is missing.

This is a property of the public data, not a defect in the pipeline. A production deployment would have credit bureau data and would need this entire analysis redone against it.

Coarsened and rounded inputs

Public HMDA deliberately coarsens several fields to protect applicant privacy:

- `debt_to_income_ratio` arrives as a band, not a number. Exact integers are retained only between 36 and 49; everything else is bucketed. The model therefore sees a coarse ordinal, and a counterfactual on this factor can only ever name a band boundary rather than a precise ratio.
- `loan_amount` and `property_value` are rounded to the midpoint of the nearest $10,000 interval. Loan-to-value ratios derived from them inherit that granularity, which places a floor on how precisely any loan-to-value counterfactual can be quoted.
- `applicant_age` is banded, so the age analysis compares bands rather than years and cannot isolate the 62-year threshold that ECOA actually names.

Derived demographic categories are reported, not observed

`derived_race` and `derived_ethnicity` are constructed by the FFIEC from what applicants reported, with their own collection artefacts. Roughly one applicant in six falls into a category that does not identify a single protected class — `Joint`, `Free Form Text Only`, or `Not Available`. Those rows are excluded from the fairness analysis rather than pooled, because a coefficient on "Race Not Available" cannot be interpreted. The counts are reported in section 7, and the exclusion is not random: non-response may itself correlate with group membership.

Missing values are imputed to make the constraints real

LightGBM does not enforce monotone constraints for rows routed through a missing-value split, and the failure is not confined to the feature that is missing — any feature carrying missing values at training time can break the guarantee for the constrained features as well. Every feature is therefore imputed at its training median before fitting. The share imputed is under 1% for every feature and is reported in section 3, but it is a real intervention: applicants with missing data are scored as though they were typical on that factor.

Clean metrics on a static dataset do not establish compliance

Everything in this document is measured on a fixed historical extract. That establishes properties of a model artifact. It does not establish that a deployed system is compliant, because deployment introduces population drift, operational overrides, distribution channel effects, and interaction with the rest of the credit process — none of which are visible here.

Counterfactuals move one factor at a time

A counterfactual holds every other input constant. Real applicant characteristics do not move independently: a smaller loan changes the loan-to-value ratio, and a higher income usually accompanies other differences. The figures are therefore a correct description of the model's decision surface and a simplification of the applicant's actual options. They should be read as "this model would have decided differently", not as financial advice.

The alternative search covers a bounded space

The less discriminatory alternative search in section 8 evaluates a finite, enumerated set of challenger models. Finding no better alternative within that space is not proof that none exists. The space searched is recorded in `artifacts/lda/grid.json` so the claim can be checked and extended rather than taken on trust.

The search also steers by one group's disparity, because a Pareto frontier needs a single scalar per axis. Other groups are reported in full but are not optimised against, and a model that improves the targeted metric may not improve theirs.

Sign-in makes the audit trail attributable, not the system secure

Manual overrides now record the authenticated user rather than a name supplied in the request, which is what makes the override trail worth keeping. That is a real improvement to the governance story and a small one to the security posture.

The authentication here is local accounts with argon2id passwords and server-side sessions. There is no federated identity, no second factor, no tenant isolation, and the demo runs over plain HTTP on a loopback address. It is adequate for a demonstration on public data and is not a model for protecting a real credit system. What is and is not implemented is set out in `docs/security.md`.

The ledger detects tampering; it does not prevent it

Append-only behaviour is enforced by SQLite triggers, and every row is chained by hash. Together these stop accidental modification and detect after-the-fact edits.

They do not stop an adversary with write access to the database file, who can drop the triggers and recompute every subsequent hash. Making the ledger genuinely tamper-*proof* requires periodically anchoring the head hash somewhere outside the system's control. That is out of scope here and is stated rather than implied.

11 Monitoring plan and revalidation triggers

What is watched after deployment, and what would force a rebuild.

What is measured, and how often

Input distribution drift

Population Stability Index on every model feature, with bin edges frozen from the training vintage and applied unchanged to each comparison period. Bands follow the conventional reading: below 0.10 stable, 0.10 to 0.25 moderate, 0.25 and above significant. Monthly in production; here it is computed between the training year and the out-of-time year.

Score distribution drift

Population Stability Index and the two-sample Kolmogorov-Smirnov statistic on the predicted probability. A score distribution can shift while every input distribution looks stable, which is why it is watched separately.

Alerting is on effect size, never on p-values. At fifty thousand observations every comparison is statistically significant, so a p-value-driven alert fires constantly and is then ignored — which is how drift monitoring becomes decoration.

Outcome and fairness monitoring

Approval rate overall and by protected group, and the controlled disparity estimate from section 7, recomputed on each monitoring window. The controlled figure is the one that matters: a raw rate can move because the applicant mix moved.

Decision integrity

The hash chain is verified from genesis on every run of `tally verify-ledger`. A break identifies the first affected row by sequence number. Any break is an incident, not a metric.

Revalidation triggers

Revalidation is required when any of the following occurs, without waiting for the scheduled cycle:

- Population Stability Index at or above 0.25 on any model feature, or on the score distribution.
- A change of 3 percentage points or more in the overall approval rate that is not explained by a deliberate policy change.
- The controlled disparity estimate for any protected group becoming statistically significant where it previously was not, or growing by more than 1 percentage point.
- Any four-fifths screening flag appearing for a group that did not previously carry one.
- A new data vintage becoming available, which changes what the out-of-time frame is.
- Any change to the feature set, the decision threshold, or the training population.
- Any break in the decision ledger hash chain.
- A change in the regulatory perimeter that affects what this documentation must contain.

Scheduled cycle

Full revalidation annually, following the release of a new HMDA vintage. Fair lending analysis quarterly. Drift and integrity checks monthly.

What monitoring cannot tell you

Every measurement described here compares the model against its own past behaviour and against a historical dataset. None of it detects a systematic problem that was present from the beginning, because the baseline itself would carry it. That is what the alternative search in section 8 and the adversarial test in the repository's test suite are for: they interrogate the model against a standard other than its own history.

12 Appendix: regulatory mapping

Each section cross-referenced to the frameworks it is written against.

These are pointers to the frameworks this document is structured against, not quotations of them, and not a claim of compliance with any of them. Current text should be verified before this mapping is relied upon.

Fed SR 11-7 / OCC 2011-12 — Supervisory Guidance on Model Risk Management

Three pillars: conceptual soundness of development, ongoing monitoring, and outcomes analysis. Applies to banking organizations supervised by the Federal Reserve and the OCC.

EU AI Act — Annex III 5(b), Article 11 / Annex IV, Article 12, Article 14

Creditworthiness assessment and credit scoring of natural persons are classified high-risk under Annex III point 5(b). Article 11 and Annex IV set technical documentation content; Article 12 covers logging; Article 14 covers human oversight. Standalone Annex III obligations were deferred to 2 December 2027 by the Digital Omnibus (Council final approval 29 June 2026). The deferral moved the deadline, not the workload — conformity assessment on a high-risk system runs closer to a year than a quarter.

ECOA / Regulation B — 12 CFR 1002

Section 1002.9 governs adverse action notification: the notice must state the specific principal reasons for the action. Model complexity does not excuse generic reasons.

FCRA — 15 U.S.C. 1681

Where a consumer report informs the decision, additional disclosure duties attach under sections 615(a) and 609(g). This system uses no consumer report data, so those disclosures are inapplicable and are deliberately omitted rather than included as boilerplate.

Section mapping

§	SECTION	SR 11-7	EU AI ACT	ECOA / REG B
1	Model purpose, intended use, and limitations of use	Conceptual soundness — model purpose and intended use	Annex IV 1(a) — general description and intended purpose	—
2	Data lineage	Conceptual soundness — data quality and relevance	Annex IV 2(d) — training data provenance, scope and characteristics	—
3	Features, transformations, and constraints	Conceptual soundness — variable selection and design	Annex IV 2(b) — design specifications and key design choices	—
4	Development methodology	Conceptual soundness — development methodology	Annex IV 2(c) — system architecture and computational resources	—
5	Performance	Outcomes analysis — performance measurement	Annex IV 2(g) — accuracy metrics and appropriateness	—
6	Stability and out-of-time validation	Outcomes analysis — stability and out-of-time validation	Annex IV 2(g) — robustness; Annex IV 5 — validation and testing	—
7	Fair lending analysis	Outcomes analysis — disparate impact and outcomes testing	Annex IV 2(f) — discriminatory-outcome mitigation measures	1002.6(a) — evaluation of applications; 1002.4(a) — general prohibition
8	Less discriminatory alternative analysis	Conceptual soundness — consideration of alternative approaches	Annex IV 2(b) — rationale and assumptions of design choices	Disparate impact — availability of a less discriminatory alternative
9	Explainability and adverse action notices	Conceptual soundness — model explainability	Article 13 — transparency; Article 14 — human oversight	1002.9(a)(2) and 1002.9(b)(2) — specific principal reasons
10	Known limitations	Conceptual soundness — documented limitations and assumptions	Annex IV 3 — foreseeable unintended outcomes and risks	—
11	Monitoring plan and revalidation triggers	Ongoing monitoring — process verification and benchmarking	Article 12 — logging; Article 72 — post-market monitoring	—
12	Appendix: regulatory mapping	Documentation — model inventory and governance record	Article 11 — technical documentation; Annex IV — full content list	—

Generated by Tallystick on 11 August 2026 at 19:25 UTC from model v002 (SHA-256 5006c9d1dac1af1b…). Every figure in this document was read from a generated artifact; none was entered by hand. Regenerating after retraining produces a document describing the new model with no manual editing.