MODEL CARD — TECHNICAL DOCUMENTATION

Illustrative Veteran Retention Logit

A weighted logistic regression that predicts a synthetic 12-month retention flag for post-9/11 veterans employed at the time of the ACS interview. Published to document the modeling workflow — feature engineering, stratified split, weighted fit, calibration — on fully public, individually de-identified data.

Author
Patrick Neil Bradley

Target
retention_12mo_synthetic (binary)

Family
Binomial GLM, logit link

Build
v0.13 · 2026-04-17

Modeled — Illustrative. The target is a synthetic label generated by a fixed-seed logistic rule. This model is a methodological demonstration, not a production retention predictor.

SECTION ONE

Summary

This model is an illustrative weighted logistic regression that predicts a synthetic 12-month retention flag for post-9/11 veterans employed at the time of the ACS interview. It exists to demonstrate an end-to-end analytics workflow — feature engineering, stratified split, weighted fit, calibration — using entirely public, individually de-identified data.

It is not a production retention predictor. The target is generated, not observed, and the sample design is not corrected for. This card is deliberately written to make those limits legible.

SECTION TWO

Target

`retention_12mo_synthetic` is a 0 / 1 flag produced by the Phase 3 pipeline using a fixed-seed logistic rule over observable cohort features. It does not come from a longitudinal employer record. Retention rates in this project are expressible as modeled-illustrative, not measured. The overall base rate is 67.9% (PWGTP-weighted) among the employed working-age cohort.

Downstream documents that cite any retention-model output must carry the **Modeled — Illustrative** label, consistent with the project convention established in Phase 3.

SECTION THREE

Cohort and universe

- Post-9/11 veterans identified via ACS PUMS variables `MIL` / `VPS` / `MLPA`.
- Ages 22–64 at interview, civilian status only (active-duty ESR 4 / 5 excluded).
- Employed at interview only (ESR ∈ {1, 2}). Retention is undefined for non-employed rows.
- Survey years 2019, 2021, 2022, 2023 pooled. 2020 intentionally omitted — ACS published a smaller experimental file for 2020 and pooling with the standard 1-year files introduces weighting inconsistencies.
- N = **108,902** unweighted rows (11.9 million PWGTP-weighted person-years) after feature engineering. 33 rows (0.03%) dropped for missing values on required features.

SECTION FOUR

Features (47 including intercept)

Family	Features
Continuous (standardized)	`age_std`, `wkhp_std`, `log_wage_std`, `unemployment_std`, `log_pc_income_std`, `attainment_std`
Demographics	`female`, `married`, race / ethnicity (reference = white non-Hispanic): `race_black`, `race_hispanic`, `race_asian`, `race_other`
Education (reference = HS / GED)	`edu_no_degree`, `edu_some_college`, `edu_bachelors`, `edu_graduate`
Disability rating (reference = not rated)	`drat_pct_0`, `drat_pct_10_20`, `drat_pct_30_40`, `drat_pct_50_60`, `drat_pct_70_100`, `drat_not_reported`
Occupation (reference = Management)	22 SOC major-group dummies from the Build v0.12 OCCP → SOC Tier-1 crosswalk
Military match	`military_skill_match` — 1 if at least one MOS maps to the person's SOC major (Build v0.12 MOS ↔ SOC match feature)

No interaction terms. No regularization. The modeling choice on this pass is transparency over accuracy.

SECTION FIVE

Split

- 70 / 15 / 15 train / validation / test.
- Seed: `20260417` (fixed at Phase 4a modeling-prep kickoff).
- Stratified on `DRAT × SOC major`, the two strongest non-demographic moderators in Phase 4a EDA.
- Split counts: train 76,145 · val 16,257 · test 16,500.

SECTION SIX

Fit

- Estimator: statsmodels GLM, family Binomial, link logit, IRLS, `max_iter=200`.
- Weights: PWGTP passed as `freq_weights`. Every coefficient and log-likelihood reflects the weighted design.
- Convergence: yes, 45 degrees of freedom on the model.

SECTION SEVEN

Metrics (held-out sets, PWGTP-weighted)

Set	AUC	Brier	Log-loss	Accuracy @ 0.5
Validation	**0.561**	**0.213**	**0.618**	**0.686**
Test	**0.566**	**0.218**	**0.627**	**0.672**

AUC is modest — by design. The synthetic target is dominated by disability, education, and state labor-market conditions, which the model captures, but the residual variance is large relative to signal. A reader who wants headline discrimination should not read this model; a reader who wants to understand how observable cohort features relate to a structured retention label should.

SECTION EIGHT

Calibration (test set, PWGTP-weighted deciles)

Decile	Mean predicted	Observed rate	Weighted n
0	0.595	0.569	152,680
1	0.627	0.633	172,040
2	0.645	0.600	175,510
3	0.667	0.674	175,600
4	0.686	0.679	176,730
5	0.696	0.684	175,140
6	0.704	0.689	183,460
7	0.711	0.722	190,240
8	0.720	0.746	191,210
9	0.736	0.726	199,720

Reading the calibration.

Observed retention rate tracks mean predicted probability within ~3 percentage points across all deciles. The model is well-calibrated even where it is not highly discriminating — i.e. it is honest about its own uncertainty.

SECTION NINE

Leading signals

Illustrative retention model — logit coefficients with 95% CIs. Grouped by feature family. Categorical references: education = HS / GED, disability = not rated, race = white non-Hispanic,

occupation = Management. Two low-sample occupation dummies (Military civilian-reported, n=11; Unknown, n=139) are excluded from the chart — their coefficients were unstable.

Ranked by absolute coefficient, dropping the two small-n occupation dummies:

- **Farming / fishing / forestry** occupation (+0.33). Small group, large positive — reference occupation is Management.
- **Disability rating 50–60%** (−0.38) and **70–100%** (−0.35). Severe-rated veterans retain meaningfully less often than non-rated peers at the same demographic and occupational position.
- **Personal care & service** (−0.14) and **food preparation & serving** (−0.11) occupations both carry retention penalties.
- **MOS ↔ SOC skill match** (+0.12). Veterans whose military occupation maps onto a civilian job in their current SOC major retain at higher rates. This is the signal the Build v0.12 match feature was engineered to capture.
- **Education: no HS diploma** (−0.12).
- **State unemployment rate** (−0.09 per SD). Tight labor markets hold cohort retention up.
- **Disability rating 0%** (+0.15). Veterans rated at 0% retain at meaningfully higher rates than the non-rated reference — consistent with the EDA finding that rated-but-unimpaired veterans fare better than the never-filed subgroup.

All seven above are significant at $p < 0.001$.

SECTION TEN

Out of scope

- **Causal claims.** The target is synthetic; no causal inference language is permissible.
- **True survey inference.** `freq_weights` reproduces the point estimate but does not produce design-correct standard errors. Confidence intervals in `coefficients.csv` are therefore directional, not survey-exact.
- **Individual-level predictions.** The model is fit to population-weighted rates; applying it to a specific veteran and treating the output as their retention probability would be inappropriate.
- **Decision support.** Nothing in this model should be used to make staffing, hiring, or benefits-administration decisions. It is a demonstration artifact.

> **If you are considering using this model for a real decision:**
>
> Don't. The target is synthetic and the sample design is not survey-corrected. This card is how you know that. Replace the target with a real observed outcome and run a fresh model card before operationalizing anything.

SECTION ELEVEN

Ethical notes

- No personally-identifying information is present. ACS PUMS has already been run through the Census Bureau's disclosure-avoidance routine; no row maps back to a named individual.
- The MOS ↔ SOC match feature is derived from the public O*NET Military Crosswalk. It is not based on any single veteran's DD-214.
- The synthetic retention target was chosen explicitly to avoid exposing any real retention data from any employer. No company, government agency, or VA system contributed to the outcome.
- Feature selection reflects the modeler's own judgment on what Phase 4a EDA surfaced as important. A different modeler would reasonably make different choices. This card documents the choices made; it does not claim they are the only defensible set.

SECTION TWELVE

Artifacts

Artifact	Purpose
`scripts/build_modeling_cohort.py`	Cohort assembly (employed-only universe)
`scripts/fit_retention_model.py`	Fit, predict, score
`scripts/build_feature_importance_chart.py`	Chart builder
`data/intermediate/modeling_cohort.parquet`	108,935 rows × 33 columns, feature-ready
`reports/phase6/coefficients.csv`	Coefficient, SE, z, p, 95% CI for every feature
`reports/phase6/metrics.json`	Headline metrics + calibration deciles
`reports/phase6/predictions_test.parquet`	Held-out test predictions for reuse
`reports/figures/phase6/feature_importance.png`	Headline chart (shown above)

SECTION THIRTEEN

What would change the story

Only two upstream deliveries materially alter this model:

- **A real observed retention outcome.** Would replace the synthetic target and allow actual causal claims about which features predict retention in this cohort.
- **O*NET work-context composites** (physical demand, schedule variability, autonomy). Currently blocked on the sandbox proxy allowlist. Once joined, the occupation-family coefficients would be decomposed into a smaller set of continuous work-environment variables, making the story about *why* some occupations retain more cleanly interpretable than 22 dummy variables.

Both are tracked in the project backlog.

FURTHER READING

Related artifacts

Full case study

The seven-chapter narrative that this model card supports. Cohort, geography, disability margin, and the model in context.

READ CASE STUDY

Executive summary (two-pager)

Thesis, four findings, federal-share tilemap, methodology note. Word document, US Letter.

DOWNLOAD .DOCX

Feature importance chart (PNG)

Headline chart from Section 9 — logit coefficients with 95% CIs.

VIEW CHART