Probabilistic Machine Learning for Parcel Forecasting at National Scale

Homecastr Technical Note

Daniel Hardesty Lewis
Founder & ML Engineer, Homecastr
March 2026

Abstract

Homecastr produces parcel-level forecast distributions, including P10, P50, and P90 trajectories at one- to five-year horizons, for residential parcels across the United States. The system combines an FT-Transformer backbone for deterministic trend estimation with a diffusion-based decoder that generates calibrated distributional samples, connected by shared inducing tokens that capture neighborhood-level spatial dynamics. Trained on over 99 million parcel-year observations from five public data sources, the model maintains calibrated 80% prediction intervals (coverage 65–95%) across forecast horizons up to five years on held-out data from both New York City and a nationwide census-tract sample. This document provides the full technical detail behind the production system described at homecastr.com/methodology: architecture, training objectives, data source characteristics, calibration diagnostics, and multi-horizon evaluation results.

1. Problem

Most consumer real-estate products present a single forecast value. A single value is easy to display, but it does not communicate downside risk, upside potential, or forecast uncertainty. For a homeowner deciding whether to sell in two years or five, the shape of the distribution matters more than the point estimate.

Homecastr produces probability distributions with P10, P50, and P90 trajectories at one- to five-year horizons for U.S. parcels. Building this required solving several problems that shaped the architecture:

- **Per-parcel calibration.** Most probabilistic forecasts are calibrated in aggregate. Making an 80% prediction interval actually contain ∼80% of outcomes for each geography and horizon, across parcels ranging from $50K rural lots to $10M urban condos, is a harder problem. Miscalibration compounds across horizons and is difficult to detect without structured evaluation infrastructure.
- **Cross-jurisdiction generalization.** Every U.S. county publishes property data in a different schema with different identifiers, vintages, and suppression rules. Building one model that generalizes across these sources required a canonical panel design that absorbs schema heterogeneity at the data layer rather than the model layer.
- **Regime sensitivity.** The model learns transition dynamics from historical data. When the current macro environment diverges from training history, as it did during the 2021 to 2022 rate shock, forecasts can degrade silently. The system needs to detect this and flag it rather than serve overconfident yet inaccurate distributions.
- **Trajectory coherence.** Forecast paths need to look like realistic extensions of a parcel's historical curve, not flat or linearized extrapolations. Independent per-horizon regressors produce trajectories that lack the curvature, momentum, and volatility present in actual price

histories. This motivated a generative architecture that samples jointly coherent multi-year paths.

2. Why Naive Approaches Fail

Before arriving at the current architecture, we explored simpler alternatives. Each failed in a specific, instructive way.

Point Forecasts Miss the Distribution

A point forecast cannot tell a homeowner whether their downside risk is 5% or 25%. Adding post-hoc Gaussian prediction intervals around a point estimate are heavy-tailed and geography-dependent. Achieving calibrated intervals requires learning the distribution directly, which motivated the move to a generative architecture.

Independent Horizons Produce Incoherent Paths

Training separate regressors for each forecast horizon ($h{=}1, h{=}2, \ldots, h{=}5$) produces trajectories that are not jointly coherent: a parcel might show 10% growth in year 2 but only 3% cumulative by year 3. A stochastic process model generates paths that respect temporal dependencies and look like realistic extensions of observed price histories.

3. System Architecture

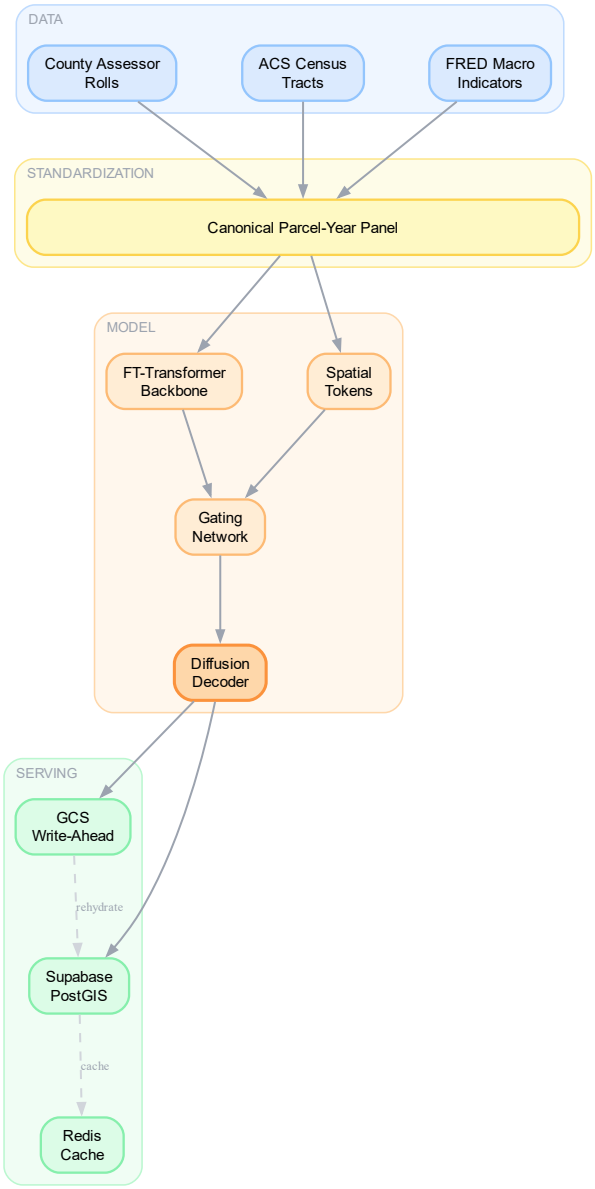

Figure 1. End-to-end system architecture. Source data flows through a canonical panel into the model stack. The FT-Transformer produces a trend estimate; the Diffusion Decoder generates calibrated distributions conditioned on both the backbone output and spatial token paths. Results stream to GCS as a write-ahead buffer before bulk upsert to Supabase.

Data Ingestion

National parcel forecasting required a canonical parcel-year schema because source systems differ in identifiers, vintages, and suppression rules. That constraint directly affected temporal alignment, model input availability, and backtesting design.

Source	Coverage	Scale	Notes
TxGIO	Texas statewide	27M parcels	Missing 2020 vintage; requires year carry-forward
Florida DOR	Florida statewide	10M parcels	String-typed appraisal values; requires explicit casting
HCAD	Harris County, TX	1.8M parcels	Richest structural features; parcel-level since 2005
NYC DOF RPAD	New York City	1.1M parcels	BBL identifiers; spatial join with MapPLUTO
ACS	US nationwide	82K tracts	Negative sentinel codes for suppressed data below −600M
FRED	Macroeconomic	11 series	Annual snapshots; intra-year movements not captured

Parcel-level features are combined with macroeconomic covariates including mortgage rates, federal funds rate, CPI, unemployment, VIX, and oil prices, sourced from FRED and joined by year.

AI-Native Interaction Layer

While traditional real estate forecasting platforms rely purely on graphical interfaces with rigid filters, Homecastr surfaces its probabilistic outputs via an **AI-Native Interaction Layer** designed for both human and agentic consumption:

- **Semantic Omnibar:** A natural language gateway that parses complex natural language queries, for example "Show me areas in Austin with >15% upside," maps them to neighborhood geometries, and instantly retrieves the relevant forecast distributions.
- **Model Context Protocol API:** Homecastr operates as a standard MCP server, exposing forecast distributions, appreciation outlooks, and comparable market data as standardized tools. This allows external reasoning models such as Claude, or custom agentic swarms, to pull raw probabilistic context directly into their prompt windows for advanced downstream analysis.

4. Model Architecture

Why a Generative Architecture Was Needed

Three limitations of simpler forecasting approaches motivated the move to a generative model:

1. Independent per-horizon regressors produce trajectory paths that are not jointly coherent. A stochastic process model generates paths that respect temporal dependencies.
2. Parametric intervals, such as Gaussian bands around a point forecast, are unlikely to capture the heavy-tailed, skewed distributions observed in realized property price changes. Learning the distribution directly avoids this assumption.
3. Neighborhood-level growth dynamics, such as gentrification and zoning changes, require latent spatial structure that tabular features alone cannot represent.

Current Production Architecture

The current model has three components, trained jointly.

FT-Transformer Backbone

A self-attention encoder over heterogeneous tabular features, including structural, locational, and macroeconomic inputs, that outputs a deterministic trend prediction $\hat{\mu} \in \mathbb{R}^H$ and a context embedding for the diffusion decoder.

Each numeric feature gets its own learned linear projection to d_{model} dimensions, alongside per-category embeddings, a history summary token, a region token, and a learnable classification-style summary token. The summary token output feeds a two-layer head that produces $\hat{\mu}$.

Inducing Tokens with Learned Persistence [?]

K=8 shared latent token paths capture neighborhood-level shocks. Each token follows an AR(1) process with learned persistence ϕ_k:

$$z_{k,t} = \phi_k \, z_{k,t-1} + \sqrt{1 - \phi_k^2} \, \eta_{k,t}, \quad \eta_{k,t} \sim \mathcal{N}(0, 1) \tag{1}$$

where $\phi_k = 0.99 \cdot \sigma(\text{logit}_k)$. The persistent parameters are constrained to the $(0, 0.99)$ interval.

A gating network computes per-parcel sparse mixing weights $\alpha_i \in \mathbb{R}^K$ using top-k selection (k=4 of 8 active) followed by softmax. The per-parcel shared driver is:

$$u_i = \sum_{k=1}^{K} \alpha_{i,k} \, z_k \tag{2}$$

In production training runs, the effective number of active tokens stabilizes near 4, confirming the sparse mixing works as intended.

Diffusion Decoder

We use a diffusion-style decoder to sample coherent multi-horizon trajectories conditioned on the backbone context embedding. The current implementation uses DDIM-style sampling (Song et al., 2020) with horizon-dependent stochasticity:

$$\eta(h) = \eta_{\text{base}} + \eta_{\text{slope}} \cdot h \tag{3}$$

with $\eta_{\text{base}} = 0.3$ and $\eta_{\text{slope}} = 0.1$. This linearly increases the noise injection at longer horizons, where epistemic uncertainty is higher and wider scenario fans are appropriate.

The diffusion operates on residuals: $r = y - \hat{\mu}$, so the backbone handles trend and the decoder handles distributional shape.

Training Objective

The model predicts annual log-growth changes ($d_{t+k} = y_{t+k} - y_{t+k-1}$) rather than absolute levels. The combined loss is:

$$\mathcal{L} = \lambda_\mu \cdot \mathcal{L}_\mu(\text{Huber}) + \lambda_\varepsilon \cdot \mathcal{L}_\varepsilon(\text{min-SNR MSE}) \tag{4}$$

The diffusion component uses min-SNR weighting (Hang et al. 2023) to prevent high-SNR timesteps from dominating. Without this, we observed that flat MSE caused the model to learn conditional means but not distributional shape, resulting in severely under-dispersed prediction intervals.

Per-horizon loss normalization averages gradients independently per horizon, preventing forecast attenuation where the model concentrates all capacity on short-term predictions.

Training Metrics

Origin	Train Rows	Final Loss	k_{eff}	σ_u
2019	362K	0.288	3.99	0.99
2021	561K	0.183	3.99	0.97
2022	761K	0.167	3.99	0.96
2023	962K	0.147	3.99	0.96
2024	1.16M	0.139	3.99	0.95
2025	5.88M	0.377	3.99	0.84

Table 1. Summary metrics from production training runs. The 2025 origin trains on the full panel of approximately 5.9 million rows and exhibits higher loss due to noise from recent regime shifts.

5. Evaluation

Temporal Holdout Protocol

All model candidates are evaluated on genuinely held-out origin years using an expanding-window protocol. We never evaluate on data that could have leaked from the training set. Macro features are lagged by one year relative to the origin to prevent look-ahead contamination.

Calibration Requirements

Each model candidate is evaluated against a structured suite of calibration diagnostics. These targets guide model development and flag regressions rather than serving as strict pass/fail gates; origins that span regime breaks are expected to fall outside target ranges. We classify calibration issues into three common cases. **Forecast attenuation**: step deltas drop to near-zero beyond the first year, producing flat trajectories. **Miscalibrated dispersion** occurs when prediction intervals are either too narrow, indicating overconfidence, or too wide, becoming uninformative, to accurately match realized outcome rates. **Mean bias** refers to systematic over- or under-prediction, typically caused by regime drift. Each case has a separate diagnostic and a corresponding remediation path.

Baseline Comparison

The calibration packet compares the model against two baselines: a persistence forecast that simply repeats the last known value, and a random-walk forecast using anchor $\times$ $(1 + \bar{g} \cdot h)$, where $\bar{g}$ is the historical mean step growth. The target is for the model to beat persistence on median absolute error for each origin-horizon combination.

Backtest Results

Each row reports the longest verifiable horizon for that origin year — the furthest point at which we can compare the model's forecast against realized values. Bold values meet horizon-adjusted

Test	What It Checks	Target	Threshold
Anchor Integrity	Forecast start matches last observed price	Median log-ratio is small	< 0.10
Interval Coverage	Outcomes fall inside 80% band	Coverage in expected range	65–95%
Horizon Scaling	Uncertainty grows at expected rate	Variance ratio near $\sqrt{h}$	±30%
Point Accuracy	Median absolute error of P50 vs. realized	MdAE within acceptable range	$< 10\%$ at h=1
Baseline Beat	Model vs. persistence forecast	Model wins on more parcels	$> 50\%$
Tail Accuracy	Extremes occur at expected rate	Each tail near nominal	$\approx 5\%$
Variance Ratio	Forecast spread matches historical	Ratio in plausible range	0.3–3.0
Distribution Match	Growth distribution resembles history	KS test does not reject	$p > 0.01$

diagnostic targets: MdAE $< 10\% \times \sqrt{h}$, Coverage 65–95%, Wins $> 50\%$. Full per-horizon results are in Appendix A.

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
NYC RPAD	2025	1	**7.0%**	**87.2%**	**56.7%**
NYC RPAD	2024	2	**8.5%**	**96.7%**	**51.8%**
NYC RPAD	2023	3	**9.8%**	**94.2%**	**63.6%**
NYC RPAD	2022	4	**12.6%**	**94.0%**	**73.6%**
NYC RPAD	2021	5	**16.1%**	**87.9%**	**59.1%**
NYC RPAD	2020	5	**12.6%**	**90.2%**	**72.0%**
NYC RPAD	2019	5	48.6%	**89.1%**	12.4%
ACS Nationwide	2023	1	**7.3%**	**71.2%**	41.7%
ACS Nationwide	2022	2	**12.7%**	**68.6%**	**87.1%**
ACS Nationwide	2021	3	30.2%	40.4%	**88.8%**
ACS Nationwide	2020	4	35.6%	46.8%	**93.2%**
ACS Nationwide	2019	5	42.8%	52.9%	**87.9%**

Table 2. Furthest-horizon backtest results for the v12sb production model. Each row shows the longest verifiable horizon for that origin. Bold values meet horizon-adjusted targets. The 2019 origins span the 2021–22 rate shock; elevated MdAE at h=5 for those vintages is expected given the regime break.

6. Serving

Inference runs on Modal using A100 GPUs with deterministic sharding and a GCS write-ahead buffer for fault tolerance. Parcel-level results are rolled into pre-materialized geographic aggregations at the ZCTA, tract, and neighborhood levels for sub-second map rendering. Feature retrieval uses Redis with tiered LRU caches.

7. Monitoring

Continuous Calibration

The calibration tests described above are not a one-time gate. As new actuals arrive, the same test suite runs against the production model to detect calibration drift. If interval coverage, tail accuracy, or horizon scaling degrades below its pass threshold, the model is flagged for retraining.

The pattern of which tests fail is diagnostic. Coverage dropping while anchor integrity holds suggests miscalibrated dispersion. Anchor integrity failing suggests mean bias from regime drift. Horizon scaling failing while short-term metrics hold suggests forecast attenuation. Each pattern maps to a different remediation path.

8. Limitations

Appraisal vs. transaction prices. The model trains on county-appraised values, not recorded sale prices. Appraised values can lag market movements, smooth over short-term volatility, and diverge from actual transaction prices in rapidly changing markets.

ACS self-reporting. Census tract demographics from the American Community Survey are self-reported survey estimates, not administrative records. They carry sampling error and non-response bias, particularly in small geographies and hard-to-reach populations.

Data coverage. Not all U.S. jurisdictions publish property assessment data in machine-readable formats. Prediction intervals are wider in data-sparse regions because the input signal is weaker.

Regime sensitivity. The model learns transition dynamics across its training history. Regime breaks, such as the interest rate shock observed between 2021 and 2022, can cause systematic bias when the current regime diverges significantly from historical patterns.

Attribution gap. Feature attributions are computed from the deterministic backbone, not the full stochastic pipeline. There is a gap between the sum of explained drivers and the calibrated trajectory.

9. Roadmap

- **Universal coverage via distillation.** This roadmap item involves a student model trained to replicate the teacher using only universally available features such as coordinates, building footprints, elevation, transit proximity, climate risk, and land use. This will enable parcel-level forecasts beyond the four jurisdictions with direct assessment data.
- **Post-hoc attribution and explainable forecast UI.** The system already computes approximate local feature attributions from the deterministic backbone using a separate post-hoc pass. The attribution step extracts feature gradients from the FT-Transformer and maps them to raw numeric variables, producing directional driver summaries such as a 3.2% increase attributed to rate cuts or a 1.1% decrease due to demand shifts. Running attribution as a separate step was a deliberate tradeoff to avoid adding latency and memory pressure to the batch inference pipeline. The next step is to surface these attributions as driver summaries in the production interface.
- **Higher-frequency macro conditioning.** The current system uses annual snapshots. Higher-frequency conditioning with decay weighting would better capture intra-year rate movements.

- **Learned heteroscedastic scale.** A learned $\sigma(x, h)$ replacing the per-horizon temperature scalar would produce geography-dependent interval widths without post-hoc tuning.
- **Causal feature integration.** Incorporating structured indicators of local policy changes, such as zoning and permitting actions, as explicit interventions would improve regime-change sensitivity.
- **Interactive scenario analysis.** This would enable users to condition forecasts on hypothetical macro scenarios, for example exploring the effect of a 200 basis point rate reduction.

References

[1] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko. *Revisiting Deep Learning Models for Tabular Data.* NeurIPS 2021. arXiv:2106.11959

[2] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, Y. W. Teh. *Set Transformer: A Framework for Attention-based Permutation-Invariant Input.* ICML 2019. arXiv:1810.00825

[3] J. Ho, A. Jain, P. Abbeel. *Denoising Diffusion Probabilistic Models.* NeurIPS 2020. arXiv:2006.11239

[4] J. Song, C. Meng, S. Ermon. *Denoising Diffusion Implicit Models.* ICLR 2021. arXiv:2010.02502

[5] T. Hang, S. Gu, C. Li, J. Bao, D. Chen, H. Hu, Z. Geng, B. Guo. *Efficient Diffusion Training via Min-SNR Weighting Strategy.* ICCV 2023. arXiv:2303.09556

[6] K. E. Case, R. J. Shiller. *The Efficiency of the Market for Single-Family Homes.* American Economic Review, 79(1):125–137, 1989. NBER w2506

[7] T. Gneiting, A. E. Raftery. *Strictly Proper Scoring Rules, Prediction, and Estimation.* JASA, 102(477):359–378, 2007. doi:10.1198/016214506000001437

[8] Florida Department of Revenue. *Property Tax Data Portal.* https://floridarevenue.com/property/Pages/DataPortal.aspx

[9] NYC Department of Finance. *Real Property Assessment Data, RPAD.* https://data.cityofnewyork.us/

[10] Texas Geographic Information Office. *StratMap Land Parcels.* https://data.geographic.texas.gov/

[11] Harris County Appraisal District. *Public Data.* https://hcad.org/

[12] U.S. Census Bureau. *American Community Survey 5-Year Estimates.* https://www.census.gov/programs-surveys/acs

[13] Federal Reserve Bank of St. Louis. *Federal Reserve Economic Data, FRED.* https://fred.stlouisfed.org/

For engineering inquiries: connect on LinkedIn

Live forecasts: homecastr.com

Multi-Horizon Backtest Results

Full backtest results for the v12sb production model across all available origin years and forecast horizons. Bold values meet the diagnostic targets from the calibration table: MdAE $< 10\% \times \sqrt{h}$, Coverage 65−95%, Wins $> 50\%$.

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
NYC RPAD	2019	1	21.2%	**89.5%**	34.3%
NYC RPAD	2019	2	34.8%	**80.7%**	6.0%
NYC RPAD	2019	3	35.8%	**85.4%**	11.6%
NYC RPAD	2019	4	39.7%	**93.8%**	10.1%
NYC RPAD	2019	5	48.6%	**89.1%**	12.4%
NYC RPAD	2020	1	**7.4%**	**91.5%**	**52.0%**
NYC RPAD	2020	2	**8.9%**	**89.2%**	**58.0%**
NYC RPAD	2020	3	**10.4%**	**90.1%**	**69.8%**
NYC RPAD	2020	4	**13.0%**	**88.0%**	**81.6%**
NYC RPAD	2020	5	**12.6%**	**90.2%**	**72.0%**
NYC RPAD	2021	1	11.9%	64.9%	29.9%
NYC RPAD	2021	2	**11.2%**	**87.4%**	50.0%
NYC RPAD	2021	3	**14.4%**	**86.5%**	**58.1%**
NYC RPAD	2021	4	**15.5%**	**86.8%**	**50.7%**
NYC RPAD	2021	5	**16.1%**	**87.9%**	**59.1%**
NYC RPAD	2022	1	**8.0%**	**94.8%**	**62.9%**
NYC RPAD	2022	2	**11.4%**	**91.8%**	**78.2%**
NYC RPAD	2022	3	**11.5%**	**94.5%**	**68.5%**
NYC RPAD	2022	4	**12.6%**	**94.0%**	**73.6%**
NYC RPAD	2023	1	**7.1%**	**91.9%**	**56.0%**
NYC RPAD	2023	2	**8.0%**	**94.4%**	**56.2%**
NYC RPAD	2023	3	**9.8%**	**94.2%**	**63.6%**
NYC RPAD	2024	1	**7.7%**	**90.8%**	37.5%
NYC RPAD	2024	2	**8.5%**	**96.7%**	**51.8%**
NYC RPAD	2025	1	**7.0%**	**87.2%**	**56.7%**

Table A1. NYC RPAD backtest results across all available origin years and forecast horizons.

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
ACS Nationwide	2019	1	**7.9%**	**86.5%**	19.7%
ACS Nationwide	2019	2	**12.2%**	**82.0%**	32.2%
ACS Nationwide	2019	3	28.9%	55.1%	31.8%
ACS Nationwide	2019	4	36.3%	48.6%	**63.2%**
ACS Nationwide	2019	5	42.8%	52.9%	**87.9%**
ACS Nationwide	2020	1	**7.2%**	**82.2%**	18.0%
ACS Nationwide	2020	2	20.5%	49.2%	**70.0%**
ACS Nationwide	2020	3	29.1%	45.7%	**75.9%**
ACS Nationwide	2020	4	35.6%	46.8%	**93.2%**
ACS Nationwide	2021	1	16.5%	34.8%	15.5%
ACS Nationwide	2021	2	21.8%	42.1%	**86.4%**
ACS Nationwide	2021	3	30.2%	40.4%	**88.8%**
ACS Nationwide	2022	1	**8.2%**	**72.3%**	29.3%
ACS Nationwide	2022	2	**12.7%**	**68.6%**	**87.1%**
ACS Nationwide	2023	1	**7.3%**	**71.2%**	41.7%

Table A2. ACS Nationwide backtest results across all available origin years and forecast horizons.