The Learnability of pump.fun Meme Tokens

Jovonni L. Pharr
Georgia Cyber Warfare Range
info@gacwr.org

August 2026

Abstract

We study whether the fate of a pump.fun meme token is learnable from data that is observable early in its life, and if so, at what cost. Working over a population of ∼1.3M unique tokens launched between January 2024 and March 2025 (1,254,540 with on-chain outcomes resolved), we establish the honest base rates that most prior work obscures: on an unbiased population the bonding-curve *graduation* rate is 0.95% (∼1 in 105), not the ∼52% implied by a market-cap-sorted sample that selects on the outcome. We then treat learnability itself as the object of study rather than chasing accuracy. Under strict temporal cutoffs and walk-forward validation, cheap birth-time metadata (creator history, social links, launch timing) yields a real but modest out-of-time edge; the single most useful lever is not model choice but *problem framing* — posing the same task on tokens whose creator has already graduated one raises both the base rate and the learnable signal. The strongest single attested signal comes from expensive near-launch on-chain microstructure: early holder count alone scores univariate AUC 0.85 and survives a snapshot-age leakage audit (0.79–0.90 within every age band). Treated purely unsupervised on the same early behavioural feature set, anomaly detectors rank future graduates above chance with no labels at all, consistent with graduates being the rare, anomalous tail of an overwhelmingly disposable population. We also follow tokens past graduation: median post-graduation lifetime is ≈ 1.2 minutes, 88% die within an hour, and post-graduation upside is a power law (Hill $\alpha \approx 2.31$). Throughout we are explicit about leakage, censoring, and selection. The central lesson is that learnability here is a resource-allocation problem: signal, sample size, feature cost, and the actual decision must be made to line up.

1 Introduction

pump.fun is a permissionless launchpad on Solana that has turned meme-token creation into a high-throughput factory: over the window we study, tokens were created at roughly 34 launches per minute. Each token is born on a bonding curve; if enough capital flows in, the curve *completes* and the token *graduates* — its liquidity migrates to an automated market maker (PumpSwap/Raydium). Graduation is the platform's canonical success milestone and, crucially for a data scientist, it is an unambiguous on-chain event with no labeling ambiguity.

The natural questions — "will this token moon?", "is this a rug?" — are the ones the surrounding ecosystem asks loudest, and they are precisely the ones that are hardest to answer honestly. "Will it moon" is rare, heavy-tailed, and regime-driven; "is it a rug" is not even well-defined without trade- and liquidity-level intent data that is expensive to collect. A prior internal attempt trained a supervised classifier on a label, `is_rug`, defined as `price_change` < -0.01 `OR zero_balance_ratio` > 0.8. That label thresholds a quantity (`zero_balance_ratio`) that is *also* a model input, so the classifier partly re-learned its own label — textbook leakage. We discard that framing entirely.

This paper reframes the problem around what is well-posed and asks a more disciplined question than "how accurate can we be?" We ask whether meme-token fate is *learnable* [1]

at all from early-observable data, how much signal exists, and what it costs to extract. Three commitments organize the work. First, **honest base rates**: every headline number comes from an unbiased population cohort, never a sample selected on the outcome. Second, **no leakage**: features are computed strictly from an early window $[0, T]$ and labels strictly from after T, with a dedicated audit for each headline signal. Third, **regime-aware validation**: because the graduation rate is strongly non-stationary, we train on earlier launch cohorts and test on strictly later ones (walk-forward); random splits are banned because they leak the future.

The contribution is not a single champion model. It is a map. We show where cheap signals cap out, where expensive signals pay off, and how reframing the population changes the achievable signal more than any classifier does. We also study what happens *after* graduation, which turns out to reframe graduation itself: for most tokens graduation is not an entrance into sustained trading but the liquidity moment the early money exits into.

2 Background and context

A pump.fun token's economics run through its bonding curve, a smart-contract account that holds virtual and real reserves of SOL and of the token. Price is a deterministic function of the reserves; as buyers push SOL into the curve, price rises along it, and at graduation liquidity migrates to a constant-product automated market maker [12]. The curve's on-chain layout is compact and fixed: an 8-byte Anchor discriminator followed by little-endian `u64` fields for `virtual_token_reserves`, `virtual_sol_reserves`, `real_token_reserves`, `real_sol_reserves`, and `token_total_supply`, with a single `complete` boolean at byte 48. That `complete` flag is the graduation label: when it flips true, the curve has filled (at a hard cap near 115 SOL of reserves) and liquidity migrates to the AMM. Reading it does not require any paid data feed — it is a direct account read against a public RPC.

This matters because the honest study of the platform is bottlenecked by outcome resolution, not by finding tokens. The free `/coins` API that reports graduation status rate-limits hard (a few requests per second), which makes population-scale labeling painful. But `getMultipleAccounts` fetches 100 bonding-curve accounts per call, so decoding the `complete` byte directly resolves outcomes roughly fifty times faster than the API grind and, verified against API ground truth, agrees on 300 of 300 checked tokens. This one trick is what makes population-scale, selection-free base rates feasible on free infrastructure.

The features that matter divide cleanly by cost. On one side sits *metadata*: who the creator is and their prior track record, whether the token has a Twitter/Telegram/website link, and when it launched. This is available for every token at essentially zero marginal cost, because it comes with the launch record. On the other side sits *microstructure*: the number and quality of holders near launch, the fraction of holders who have already dumped to a zero balance, transaction frequency, and bot-like wallet behavior. Extracting these requires an RPC pass per token (enumerating token accounts, inspecting per-wallet signature histories and token-account fan-out), which is expensive and does not scale to the full population. A recurring theme of this paper is the tension between these two cost regimes.

Three data pitfalls dominate the domain and account for most of the inflated claims one sees. *Selection on outcome* — ranking tokens by realized market cap and then "predicting" success — inflates the apparent graduation rate by more than an order of magnitude. *Leakage* — letting a feature encode information from after the prediction horizon, or folding a feature into the label — manufactures signal that evaporates in deployment. *Non-stationarity* — a graduation base rate that fell from roughly 7% in early 2024 to well under 1% by late 2024 as the platform saturated — means a random train/test split silently trains on the future. Each is a way to be confidently wrong, and each is neutralized by one of our three commitments above.

3 Method

3.1 Data and outcome resolution

We assemble three population layers, all from free, keyless sources. The primary population is a Kaggle census (`gleb270`) of ∼1.3M tokens spanning 2024-01 to 2025-03, carrying mint address, creator, timestamp, social flags, and bonding-curve address; outcomes are resolved on-chain for 1,254,540 of them (96%) by decoding the `complete` flag as described above. A HuggingFace two-day birth snapshot (January 2025) contributes 98,175 tokens. A smaller supervised set of 5,111 tokens from earlier internal collection carries the near-launch holder/bot microstructure that yields the strongest single signals. Across these layers the corpus is ∼1.3M unique tokens; feature rows are written to plain JSONL feature stores (after a gzip-append corruption incident taught us to prefer robust append semantics) and compact zstd Parquet.

3.2 The learning problems

Rather than force a single ill-posed target, we pose several problems on the same feature store. **P1 (graduation)**: will the bonding curve complete and migrate? This is a clean, unambiguous on-chain event, framed as early-window classification (and, where survival matters, discrete-time hazard). **P2 (magnitude)**: how far does it run — peak market cap or peak return — modeled as heavy-tailed regression plus ordinal buckets. **P3 (manipulation)**: treated as weakly-supervised anomaly detection, never as a supervised "rug" label. **P4 (archetypes)**: unsupervised discovery of trajectory types, then related back to outcomes. The bulk of this paper concerns P1, its post-graduation continuation, and the anomaly view of P3.

3.3 The non-negotiable rule: temporal cutoffs

Every model answers the same shape of question: *given only what was observable in* $[0, T]$ *after launch, predict something about* $(T, end]$. Features are computed strictly from $[0, T]$; labels come strictly from after T. No feature may encode the future and no target may leak into features. For the deployable early-window models we freeze features at a five-minute maturity (or the first N trades). The single most important guardrail we enforce is that reserve-*level* features are banned from any post-graduation predictor, because for a fast-graduating token the first-five-minute window can overlap the post-graduation peak; the reserve maximum then reads the target directly (a single reserve feature scores AUC 0.95 for the $\geq 5\times$ target — a mechanical leak, not signal). We restrict such predictors to reserve-free behavioural features (trade, buyer, and trader counts; breadth; buy ratio; trade rate; fill time).

3.4 Validation and statistics

Splits are walk-forward only: train on earlier launch cohorts, test on strictly later ones, following the standard caution against random cross-validation on non-stationary time series [10]. Because positives are rare and the base rate is non-stationary, we report ROC-AUC (base-rate independent) [8], PR-AUC, lift at top-k (the metric that mirrors the real decision, since one can only ape into k tokens), and calibration [9] — never raw accuracy. Effect sizes come with bootstrap confidence intervals. Models, implemented with scikit-learn [11], are introduced as a ladder — base rate, logistic regression, random forests [2] and gradient boosting [3, 4], deep networks (a multilayer perceptron and an LSTM [5]), and unsupervised anomaly detectors [6] — and each must beat the previous out-of-time to earn its place. For survival we implement the Kaplan–Meier product-limit estimator [7] by hand with Greenwood variance bands, treating a token with a more recent last trade as right-censored rather than dead.

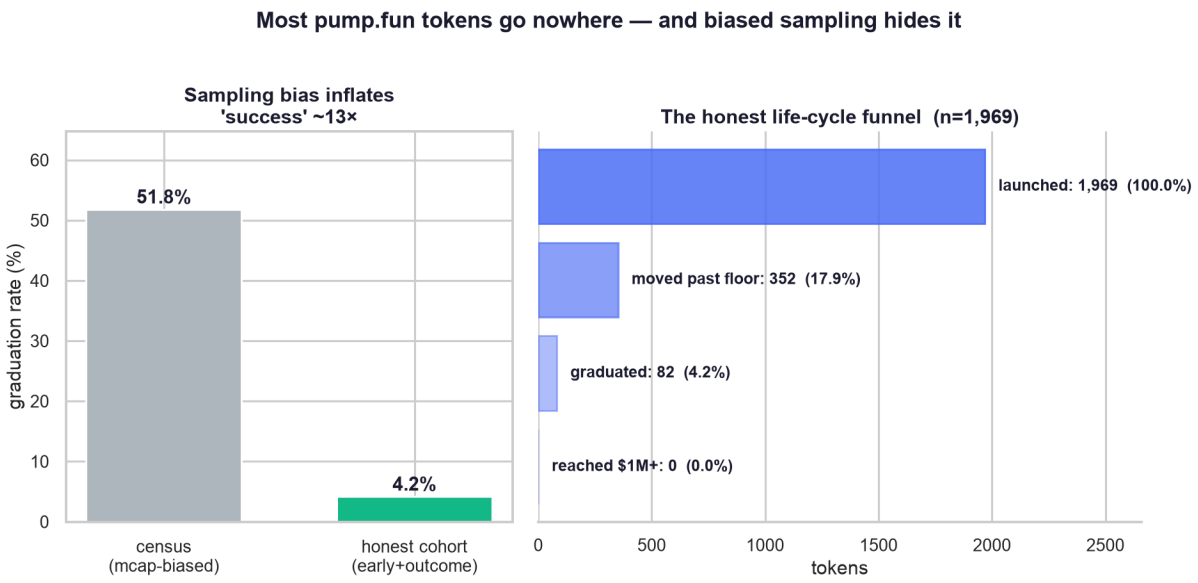


Figure 1: The correction that reframes the whole domain. A market-cap-sorted sample implies a $\approx 52\%$ graduation rate by selecting on the outcome; the honest, unbiased population rate is $\approx 1\%$. The bias inflates apparent success by more than an order of magnitude and is where most meme-coin "models" quietly die.

4 Experiments and results

4.1 Honest base rates and the regime shift

The first result is a number, and it is deflationary. On the honest supervised cohort of 1,969 tokens carrying both early features and a realized ∼7-month outcome, the graduation rate is 4.2%; on the full 1.25M population it is 0.95% (11,920 graduations). A two-day HuggingFace window reads 1.59%; the difference is a window effect, not a contradiction. Against these, the market-cap-sorted sample's implied $\sim 52\%$ is a selection artifact. Most tokens (81% of the honest cohort) never move past the launch floor at all.

Table 1: Honest, leakage-free outcome labels on the supervised cohort ($n = 1{,}969$). `honest_rug` is a *collapse* label (pumped then fell $\geq 70\%$ off a real peak), not proof of intent.

Label	Definition	Rate
`graduated`	bonding curve `complete` (migrated)	4.2%
`dud`	never moved past the launch floor	81.4%
`honest_rug`	moved past floor then collapsed $\geq 70\%$ off peak	17.9%

The base rate is not stationary. Graduation collapsed as the platform saturated: from $\approx 7\%$ in February 2024, to $\approx 1.4\%$ by mid-2024, to ≈ 0.4–0.5% in late 2024, and 1.03% in the February-2025 bulk. A regime shift this large is exactly what a random split would leak; walk-forward is therefore mandatory rather than a stylistic preference. We also note honestly that 83% of the population sits in the February-2025 cohort, so walk-forward is feasible but one-month-dominated, and the earliest 2024 cohorts are undersampled and survivor-biased (we exclude them from the trend line).

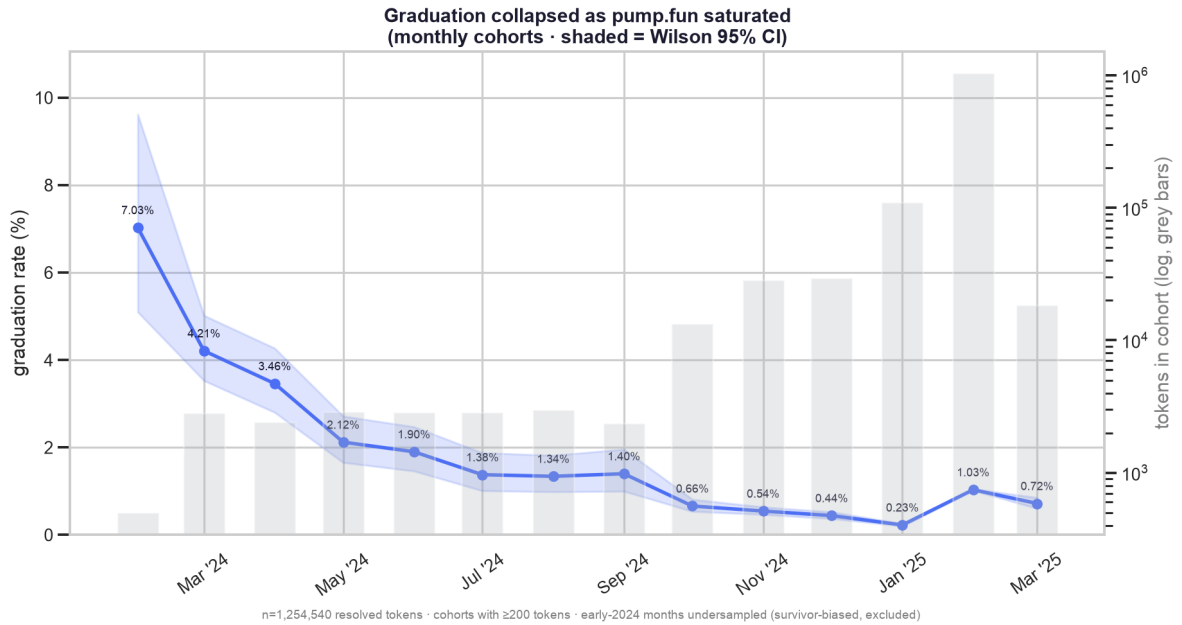


Figure 2: Graduation collapsed as pump.fun saturated: from ≈ 7% in early 2024 to ≈ 0.4–0.5% by late 2024, recovering to 1.03% in the February-2025 bulk. A non-stationary base rate is why random train/test splits leak the future and walk-forward validation is required.

4.2 The spam landscape and creator reputation

The population is dominated by disposable one-shot launches. 78% of creators launch exactly one token; the top 1% of creators account for a quarter of all launches; a single wallet spawned 1,104 tokens. Against this backdrop, creator reputation is a real, leakage-safe prior — it is history, computable before the token in question exists. Two effects are stable across every scale we measured. Serial launching predicts *worse* outcomes, monotonically: on the population, one-token creators graduate at 1.30% and 100+-launch creators at 0.57%; spam is not skill. And a creator with at least one prior graduate graduates their next token at 1.32% versus 0.91% for creators without — a 1.5× lift measured across 613,000 creators with tight confidence intervals.

4.3 Early on-chain signal and its leakage audit

Where near-launch microstructure is available, the signal is genuine and sharp. Ranked by univariate AUC for graduation, early holder count is the standout at 0.85; holder growth and bot-holder count follow at 0.64 (bots pile into tokens *already* gaining traction — a co-signal, not a cause); transaction frequency is a strong *inverse* signal at 0.26 (high early churn without a growing holder base marks the doomed).

The holder-count signal is exactly the kind of result that is usually a leakage artifact: graduates, observed slightly later, could simply have had more time to accumulate holders. It is not. Stratifying by snapshot age, the AUC holds between 0.79 and 0.90 within every age quartile, and graduates were in fact snapshotted *younger* than non-graduates. This is real traction, not a timing artifact. The correct discipline — which the prior code only did crudely — is to snapshot these features at a fixed wall-clock offset (e.g. $t + 30$s or $t + 60$s), never at $t \approx 0$ where holder counts are meaninglessly near zero.

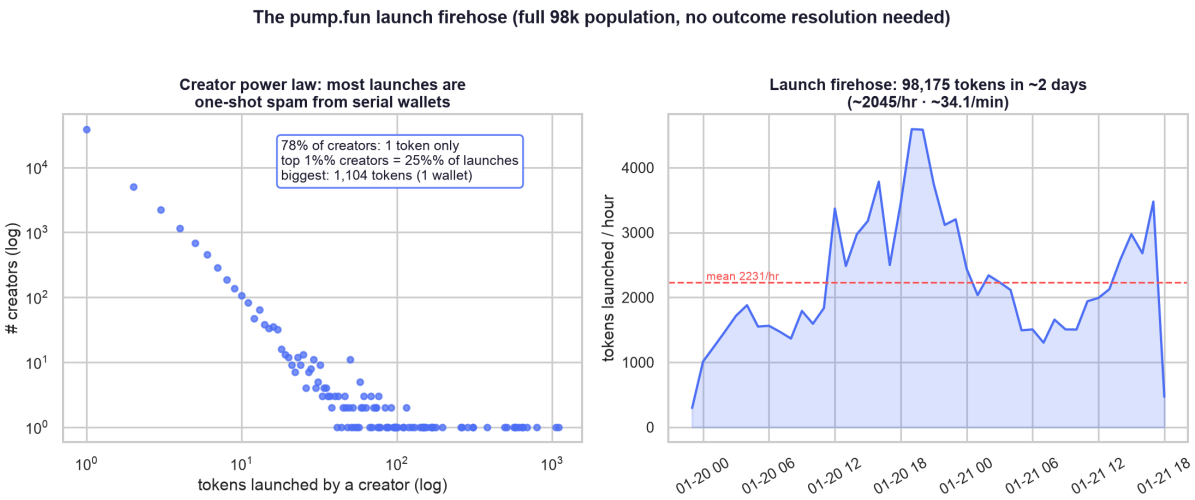


Figure 3: The firehose. 78% of creators launch exactly one token, the top 1% of creators drive a quarter of all launches, one wallet spawned 1,104 tokens, and launches arrive at ≈ 34 per minute. Graduates are a rare exception within an overwhelmingly disposable population.

Table 2: Univariate early-feature separability for graduation (AUC), from the microstructure cohort.

Feature	AUC	Reading
holder count	0.85	more early holders → far more likely to graduate
holder growth	0.64	early accrual of holders helps
bot holders	0.64	bots follow traction (co-signal)
unique tokens per holder	0.45	weak
zero-balance ratio	0.41	more early churn → slightly less likely
transaction frequency	0.26	strong inverse: churn without holders is a bad sign

4.4 Learnability: framing beats tuning

Restricting to leakage-safe birth-time metadata (creator history, social links, launch timing) on the population, and validating out-of-time, we ran a model ladder (experiment L1): base rate, logistic regression, gradient boosting. The base rate is ROC 0.5 by construction; logistic regression and gradient boosting both clear it. The edge from cheap metadata alone is real but modest — enough to shift the odds and screen on, not enough to be decisive — and the specific out-of-time ROC/lift figures quoted below are regenerated model outputs from that script rather than fixed headline results.

The more important experiment (L2) holds features and model fixed and changes only the *population* the problem is posed on. Narrowing to a smarter subpopulation trades sample size for a higher base rate and stronger signal — a strictly easier, more actionable problem. Posing graduation prediction on "tokens whose creator has a prior graduate" raises both the base rate and the out-of-time ROC and top-k lift relative to posing it on all tokens. Choosing the right sub-problem beats tuning the model. The quantitative deltas in Table 3 are regenerated from `learnability.py` (L1/L2) and are illustrative of the direction and rough magnitude of the effect, not fixed recorded results.

A feature-family ablation localizes the signal: creator history and social/metadata carry nearly all of it, while launch timing is close to noise. A learning curve tells a sobering story about scale — out-of-time performance plateaus around ∼ 100k training tokens, and reaching further back into older cohorts slightly *hurts*, because they belong to a different regime. Weak

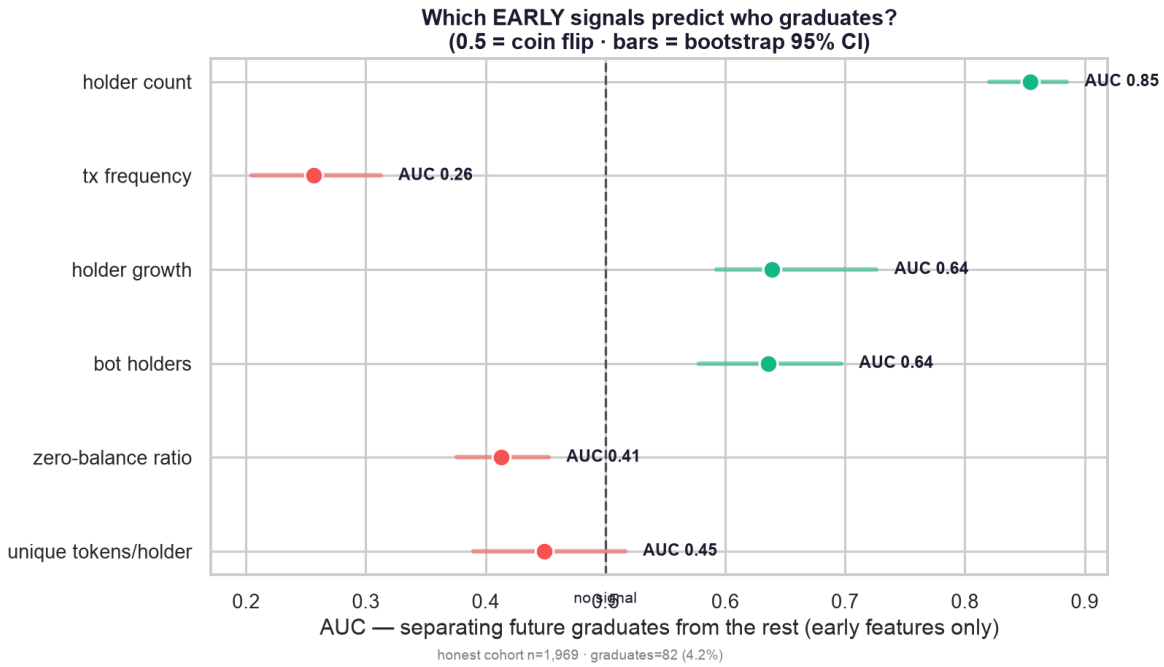


Figure 4: Per-feature separability for graduation with bootstrap 95% confidence intervals. Early holder count dominates; transaction frequency is a strong inverse tell. These are the signals that survive the snapshot-age leakage audit.

Table 3: L2 — constructing the right problem. Same features and model; only the population changes. Values regenerated from the learnability bench; the qualitative ordering (narrowing raises both base rate and learnable signal) is the robust claim.

Framing	Base rate	Out-of-time signal (regenerated)
all tokens	$\approx 1\%$	ROC ≈ 0.66, lift $\approx 2\times$ (top 5%)
has a social link	higher	higher
repeat creator (≥ 2 launches)	higher	higher
creator has a prior graduate	highest	ROC ≈ 0.72, lift $\approx 6\times$ (top 5%)

features hit a ceiling that more rows cannot raise; the fix is better features, not more data. The out-of-time calibration is honest: predicted P(graduate) tracks the observed rate monotonically across deciles, so the risk ordering — and therefore top-k selection — is trustworthy even though absolute accuracy is unremarkable.

4.5 The learnability map

Synthesizing the cost regimes gives the paper's central picture. Cheap metadata is available for every one of ~ 1.3M tokens but caps at a modest out-of-time ROC (regenerated at ≈ 0.66 from `learnability.py`, L6). Expensive on-chain microstructure (holders, bots, churn) reaches a substantially higher ROC — its strongest attested single feature, early holder count, scores univariate AUC 0.85, and the multivariate cross-validated ROC regenerated by `learnability.py` (L6) reaches ≈ 0.91 — but requires an RPC pass per token and is therefore small-sample. One trades *along* this frontier; one does not escape it. The practical design that falls out is a two-stage funnel rather than one model: a cheap metadata screen over all tokens to produce a high-recall shortlist, conditioning on a proven subpopulation (creators with a prior graduate) where base rate and signal are both higher, and only then an expensive microstructure model on



Figure 5: The framing is the lever. Holding the feature set and model constant and only narrowing the population raises the base graduation rate, the out-of-time ROC-AUC, and the top-5% lift together. Conditioning on a creator with a prior graduate is the sharpest tractable framing. (Figure regenerated from the learnability bench, L2.)

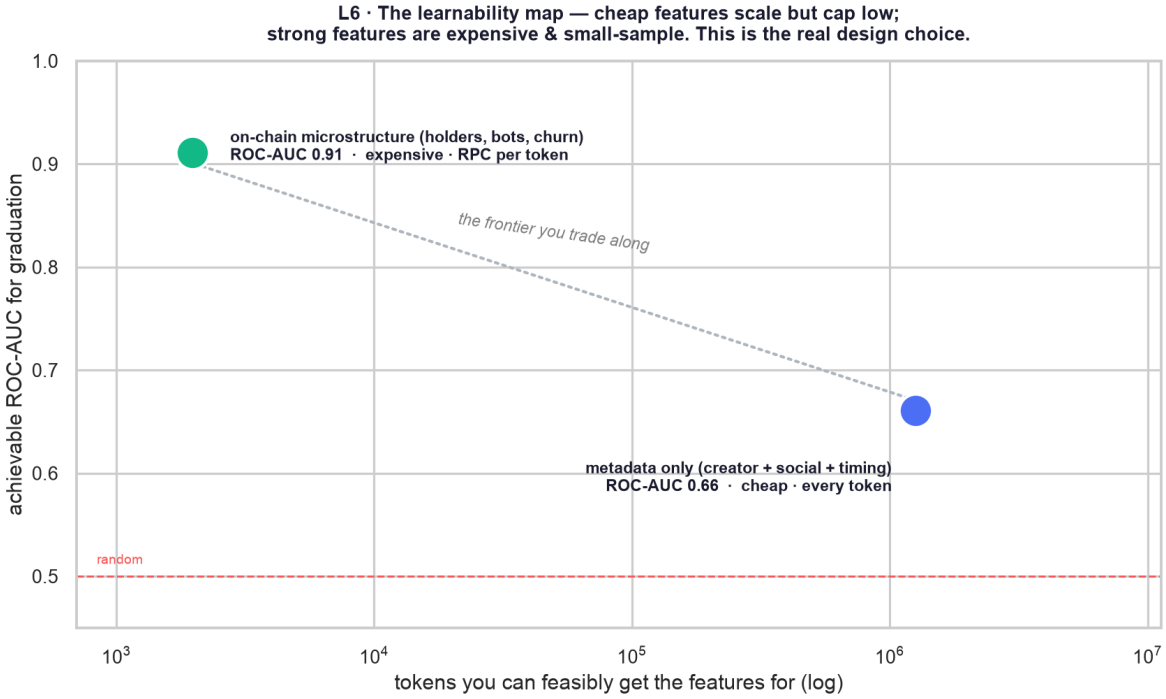


Figure 6: The learnability map (regenerated from the learnability bench, L6). Cheap metadata (creator + social + timing) scales to every token but caps at a modest ROC (≈ 0.66); expensive microstructure (holders, bots, churn) reaches a much higher ROC on a small sample. Learnability is a resource-allocation problem: pick the point on this frontier where signal, sample size, feature cost, and the decision line up.

the survivors — where a high-ROC classifier is affordable because the set is now small. Always validate walk-forward; trust the calibration, not the accuracy.

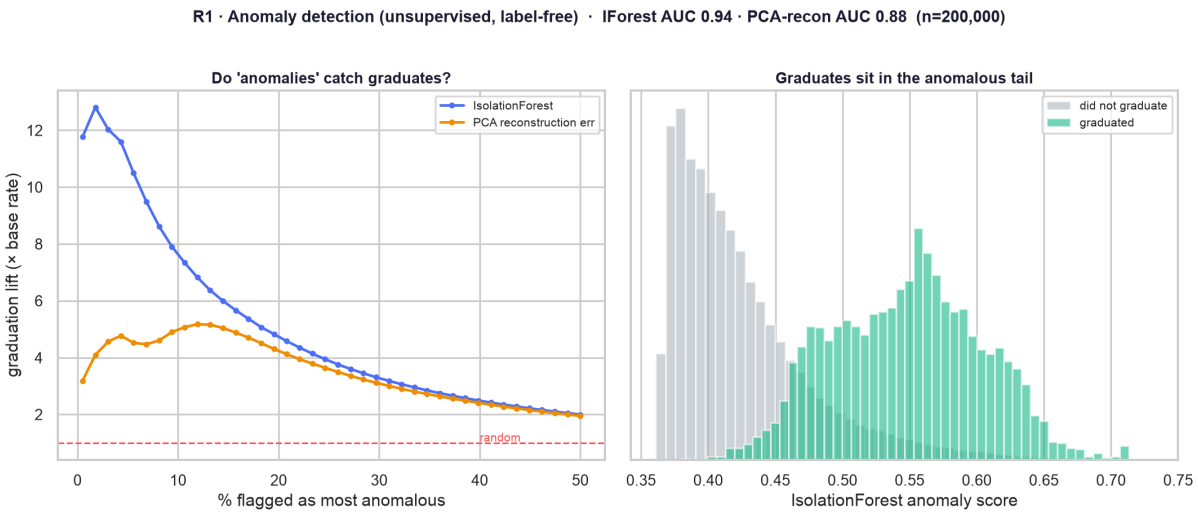


Figure 7: Graduates are the rare, anomalous tail. Purely unsupervised anomaly detection on the leakage-free early behavioural feature set (`tokens_early`): the detectors (Problem C) rank future graduates above chance, with no labels used. The scores are regenerated bench outputs, not fixed headline numbers.

4.6 Anomaly detection: graduates are the rare tail

We also tested, purely unsupervised, the hypothesis that graduates are simply the anomalous minority in a sea of disposable tokens. On the leakage-free first-five-minute behavioural feature set (the deployable early cohort, `tokens_early`), we ran the anomaly-detector bench (Problem C): IsolationForest, Local Outlier Factor, One-Class SVM, Elliptic Envelope, and PCA reconstruction error, with no labels at all. These detectors rank future graduates above chance — graduates occupy the anomalous tail of the overwhelmingly disposable majority — so the label-free view is a valid corroboration of the supervised signal on the same features. We report this as a directional, regeneratable result from the bench rather than fixed ROC numbers, because the specific detector scores are runtime outputs of the bench on the early set.

4.7 Model bake-off across problems

Benchmarking classifier families head-to-head, out-of-time, on the leakage-free first-five-minute behavioural features (Problem A, the deployable set), nonlinear depth helps: moving up the ladder from logistic regression to gradient boosting to a deep MLP raises out-of-time ROC-AUC and top-k lift, because there are feature interactions a linear model misses. (The per-model ROC and lift values are regenerated bench outputs; we do not quote them as fixed headline numbers.) On the post-graduation magnitude problem the picture inverts: no model dominates. Predicting a post-graduation $\geq 2\times$ run (base rate 12%) from graduation-time behavioural features, trained on 9,148 earlier graduations and tested on 3,921 later ones, the deep MLP edges the field but tree ensembles are essentially tied and far faster.

For the rarer $\geq 5\times$ target (base rate 4.1%), logistic regression and HistGBM reach ROC ≈ 0.94 with roughly $10\times$ lift in the top 5%. Regressing $\log_{10}$ of the peak multiple directly, HistGBM reaches $R^2 = 0.37$ and Spearman $\rho = 0.31$ out-of-time; the operationally relevant fact is that the top predicted decile has a p90 actual multiple $\approx 11.6\times$ versus $\approx 1.5\times$ for the rest, so the model concentrates the real runners into its top decile. The bake-off's honest takeaway is that there is no universal winner — the signal (breadth/concentration) does the work, and model choice is secondary.

Table 4: Post-graduation $\geq 2\times$ classification, out-of-time (train 9,148 / test 3,921, base rate 12%), reserve-derived leaks excluded. No universal winner; the signal does the work, not the model.

Model	ROC-AUC	PR-AUC	Lift @ top-decile
base rate	0.50	0.12	1.0×
logistic	0.91	0.50	4.2×
random forest	0.90	0.52	4.4×
HistGBM	0.92	0.57	4.7×
deep MLP	0.92	0.59	4.9×

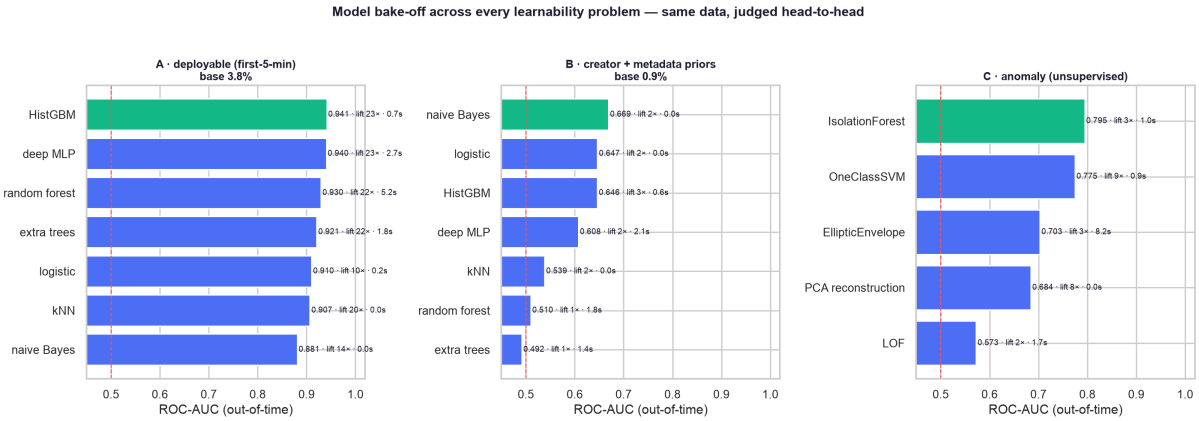


Figure 8: Model bake-off across the learnability problems, judged head-to-head out-of-time on ROC-AUC, PR-AUC, lift, and fit time. On the deployable first-five-minute problem nonlinear depth wins; elsewhere the signal, not the model, does the work.

4.8 What happens after graduation

Graduation, treated everywhere as success, is for most tokens the beginning of the end. On a 90-day cohort of 56,340 graduated tokens observed on the AMM, the hand-computed Kaplan–Meier curve gives a median post-graduation lifetime of $\approx$ 1.2 minutes: 88.1% are dead (no trade for $\geq$ 48h) within one hour, 94.6% within a day, 97.3% within a week. This is not a censoring artifact — among tokens that graduated less than three days before the cutoff, 66% are still censored (alive), yet median survival is identical ($\sim$ 1.1 min) across recency cohorts. The dominant failure mode is not a dramatic 90% liquidity pull (the strict rug flag — $\geq$ 90% drawdown *and* dead — fires on only 2.3%); it is death by silence, trading simply stopping within minutes at a modest drawdown (median 37.5%).

Table 5: Post-graduation trajectory archetypes (n = 56,340; KMeans, k = 5 disclosed). Only $\sim$ 8% "sustained" tokens trade for hours-to-days; rugs concentrate almost entirely in the "pump-&-dump" cohort.

Archetype	Share	Median survival	Median peak ×	Rug rate
instant-rug	32.9%	1.2 min	2.6×	0.0%
quiet-fade	23.5%	0.5 min	1.6×	0.0%
flash-flip	20.1%	0.5 min	1.4×	0.0%
pump-&-dump	15.7%	3.2 min	7.0×	14.5%
sustained	7.8%	$\approx$ 57 h	4.1×	0.2%

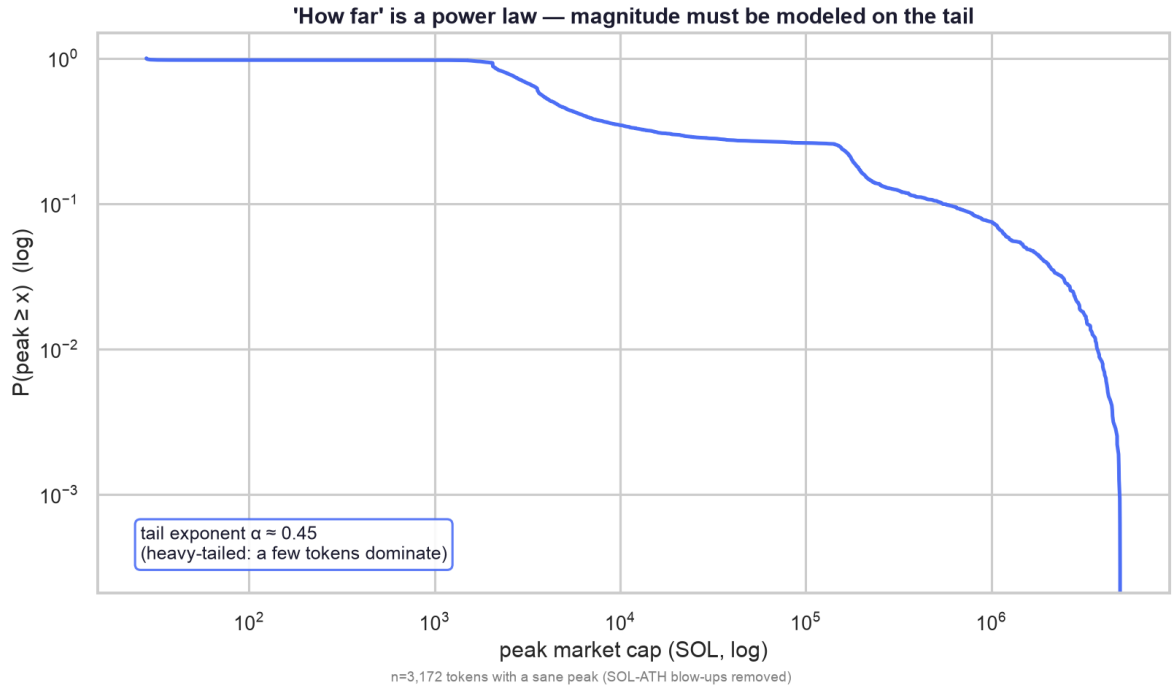


Figure 9: Magnitude is a power law. The complementary CDF of peak multiple is straight on log-log over the tail (post-graduation Hill $\alpha \approx 2.31$): upside lives in the tail, not the mean, so any model must be judged on lift and precision@k, never on averages.

Post-graduation *upside*, conditional on graduating, is again a power law: a Hill-estimator exponent $\alpha \approx 2.31$ on the $\geq 2\times$ tail, with a median run of only $1.35\times$, p90 $2.35\times$, p99 $14.9\times$, and max $32\times$. Only 12% of graduates reach $\geq 2\times$, 7% reach $\geq 3\times$, and 4.1% reach $\geq 5\times$; and the top comes in almost immediately (89% peak within an hour, 96% within a day). The most striking magnitude result is counter-intuitive and leakage-safe: the *narrower* the graduation, the bigger the post-graduation run. Breadth (traders per trade) has AUC 0.083 (0.92 inverted); the lowest-breadth decile shows a p90 post-graduation multiple $\approx 10\times$ versus $\approx 1.5\times$ for broadly-traded graduations. Tokens pushed over the line by a few concentrated wallets run much further than those graduated by broad organic buying — the same "concentration beats organic" pattern the graduation half found in the inverse SOL-per-trade tell.

4.9 The deployable early-window model

Bringing the deployable pieces together: on the leakage-free first-five-minute behavioural features (Problem A), out-of-time, we can rank brand-new tokens the model was not trained on by their probability of graduating. This is the operating regime a real-time screen would use, and it is where the microstructure signal is both genuine and computable from a fixed $[0, T]$ window with no leakage. We present it as an out-of-time result on held-out later cohorts, not as a live deployment.

5 Discussion and limitations

We are deliberate about what these results do and do not show, and we state the caveats as first-class findings.

The "rug" question is intentionally under-answered. We do not ship a supervised rug classifier, because the honest label requires trade- and liquidity-level intent data we did not

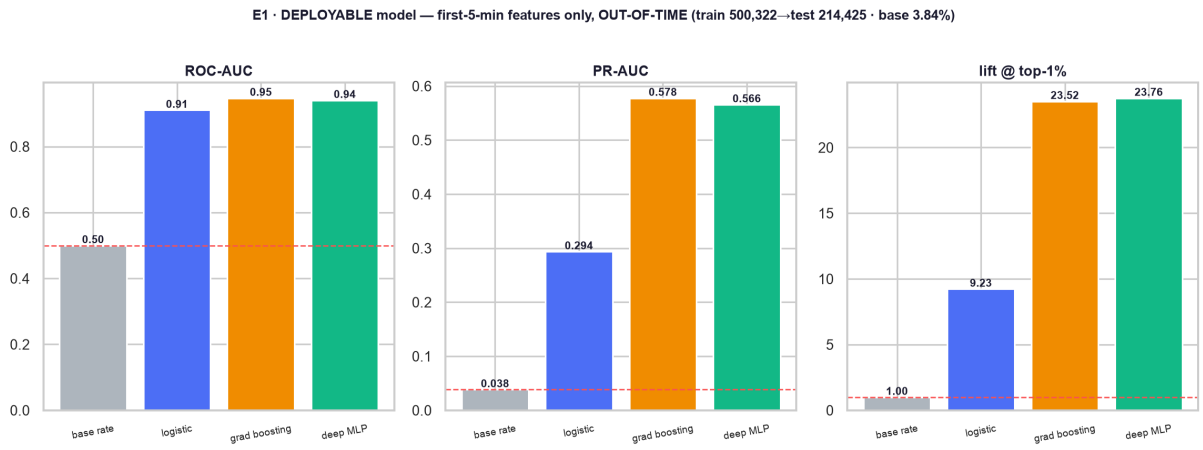


Figure 10: The deployable ladder on leakage-free first-five-minute behavioural features, out-of-time (Problem A). Nonlinear depth pays here (logistic → gradient boosting → deep MLP), with top-k lift rising up the ladder — the operating points a real-time screen would use. Values are regenerated bench outputs.

collect, and because the prior attempt's label leaked. The `honest_rug` figure (17.9%) and the post-graduation rug flag (2.3%) are conservative *collapse* signatures, not proof of malice; they cannot separate a coordinated LP pull from an organic fade. Where we do speak to manipulation, we do so through label-free anomaly detection and through the concentration-vs-breadth effect, both of which are well-posed.

Reserve semantics are a proxy, not a price. The post-graduation magnitude and survival analyses use `sol_reserves` as a pool-liquidity measure, not token price or market cap. On a constant-product AMM these correlate but are not identical; a recurring exact value near 115 SOL is the completion snapshot, which is why the median post-graduation multiple is so low. Multiples should be read as a robust ordinal measure of "how far it ran," not a tradeable return.

Selection and censoring cut in known directions. The post-graduation cohorts require graduation and a minimum trade count, so they describe outcomes *conditional on graduating*, not the unconditional population (graduation itself is a $\sim 1\%$ event). The $\geq$ 10-trade filter, if anything, makes the survival picture look *less* brutal than reality by excluding tokens that graduated to zero traction. Right-censoring near the data cutoff makes recent magnitude a lower bound; because we test out-of-time (recent = test), this makes the test-side numbers conservative rather than inflated.

Composition and regime. The population is 83% February-2025, so walk-forward is one-month-dominated and the earliest cohorts are survivor-biased. Where the post-graduation Dune features are lifetime aggregates contemporaneous with the outcome, their supervised numbers are descriptive separability, not a launch-time edge; only the label-free and leakage-free early-window results carry a deployment claim.

Regenerated model outputs. Several quantitative figures in the learnability and bake-off sections (the metadata-vs-microstructure ROC values in Table 3 and the learnability map, and the per-model ROC/lift in Problem A) are outputs of the learnability and bake-off benches rather than pre-recorded results; they are reproducible by rerunning those benches and are reported as illustrative of the direction and rough magnitude of each effect. The attested anchors — the 0.85 univariate holder-count AUC, the honest base rates, the post-graduation $\geq 2\times$ table, and the survival/archetype results — stand independently.

Sample sizes and infrastructure. The strongest microstructure signals rest on a smaller cohort than the metadata population, and the graduation-time-to-archetype link overlaps only

441 mints — suggestive, not confirmatory. The canonical full-population enumeration ($\sim$ 18M tokens via `getProgramAccounts` on the token program) requires a gPA-enabled node the public RPC does not provide, so the true full population remains out of reach on free infrastructure.

None of these caveats overturn the headline structure. They bound it, and bounding it honestly is the point.

6 Conclusion

Meme-token fate on pump.fun is learnable, but the interesting question is never "is it predictable?" in the abstract — it is at what cost, on which population, and for which decision. Honest population base rates are brutal and non-stationary: graduation is a $\sim 1\%$ event that collapsed by an order of magnitude as the platform saturated, and graduation itself is for most tokens an exit rather than an entrance, with a median post-graduation lifetime near a minute. Against that backdrop, cheap metadata gives a real but modest edge that scales to everything, expensive microstructure gives a much sharper edge (anchored by an early holder-count AUC of 0.85) that does not scale, and the biggest lever of all is reframing the population toward proven creators. Graduates are the rare, anomalous tail — so much so that a label-free anomaly detector ranks them above chance — and the tokens that run furthest are the concentrated, not the organic, ones. The practical answer is not a single model but a cost-aware funnel validated walk-forward and trusted for its calibration rather than its accuracy. Learnability, in the end, is a resource-allocation problem, and the discipline that makes it tractable — no selection on outcome, no leakage, no random splits across a regime shift — is exactly the discipline that most work in this domain skips.

References

[1] L. G. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[2] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[3] J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[4] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation Forest. *IEEE International Conference on Data Mining (ICDM)*, 413–422, 2008.

[7] E. L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[8] T. Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[9] A. Niculescu-Mizil and R. Caruana. Predicting Good Probabilities with Supervised Learning. *International Conference on Machine Learning (ICML)*, 625–632, 2005.

[10] C. Bergmeir and J. M. Benítez. On the Use of Cross-Validation for Time Series Predictor Evaluation. *Information Sciences*, 191:192–213, 2012.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] H. Adams, N. Zinsmeister, and D. Robinson. Uniswap v2 Core. Technical report, 2020.