Leakage-Free Out-of-Time Prediction of Consumer Loan Default

A Technical Report on 36-Month LendingClub Loans (2007–2018)

Training on 2007–2014 Vintages, Evaluating Out-of-Time on 2015

July 12, 2026

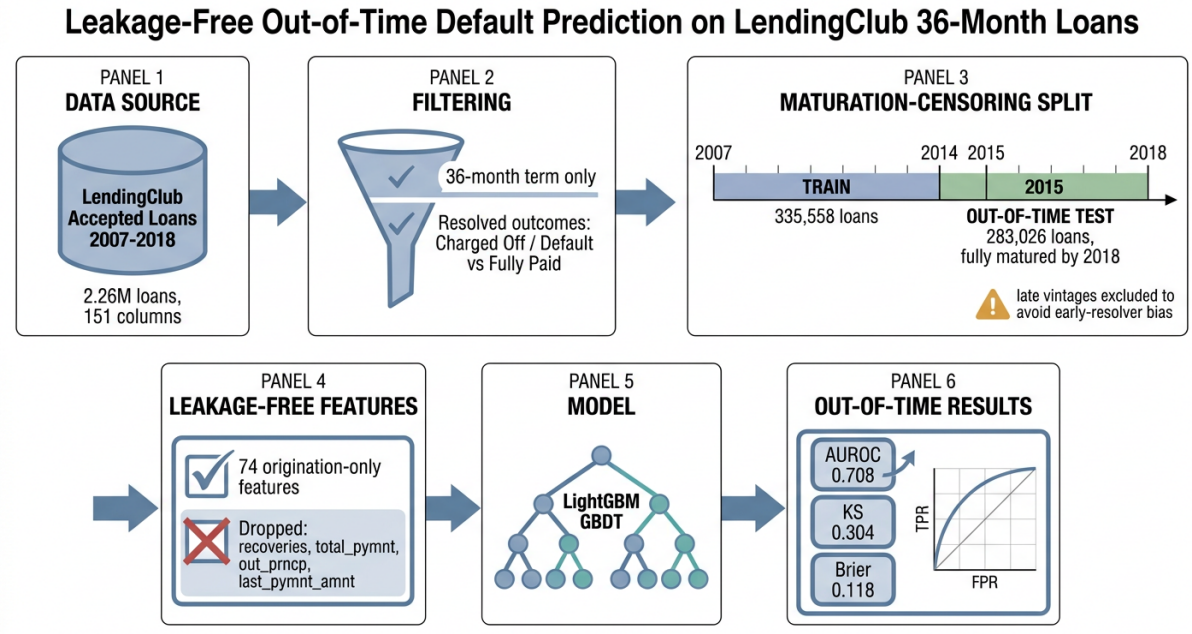


Figure 1: **Graphical abstract.** End-to-end pipeline: the full public LendingClub accepted-loans dataset (2007–2018, ≈2.26M loans, 151 columns) is filtered to 36-month resolved loans, split on a maturation-aware temporal boundary (train 2007–2014, out-of-time test 2015), reduced to 74 origination-only features (all post-origination fields removed to prevent leakage), and modelled with a LightGBM gradient-boosted decision-tree classifier. The 2015 out-of-time test yields AUROC 0.708, Kolmogorov–Smirnov 0.304, and Brier score 0.118.

1 Executive Summary

This report documents the construction and out-of-time (OOT) validation of a credit-risk model that predicts the probability that a 36-month LendingClub loan will default, using only information available at the moment of loan origination. The work was designed around a single, overriding methodological principle: the reported performance must reflect what a lender could genuinely have achieved at underwriting time, with no contamination from information that only becomes available after a loan is funded, and no distortion from the fact that recently issued loans have not yet had time to reach their contractual maturity.

Two structural pitfalls dominate naïve analyses of this dataset. The first is *target leakage*: the public LendingClub file mixes origination attributes (loan amount, interest rate, borrower income, credit-bureau summaries) with dozens of fields that are updated throughout the life of

the loan (cumulative payments, recoveries, outstanding principal, hardship and settlement flags). A model that consumes these post-origination fields can trivially "predict" default because those fields effectively encode the outcome; such a model is useless in deployment and its inflated accuracy is an artefact of leakage [17]. The second is *maturation censoring*: a 36-month loan issued late in the sample (2016–2018) has not had time to mature by the December 2018 data snapshot, so its final status is still "Current". Filtering such late vintages to "resolved only" does not yield a clean sample — it retains a biased subset of *early resolvers* (predominantly early charge-offs and early prepayers), which systematically misrepresents the true default rate.

We address both pitfalls directly. Every post-origination field is enumerated and dropped; the feature set is restricted to 74 origination-time predictors. The temporal split is chosen so that the evaluation cohort is fully matured: a 36-month loan issued in 2015 reaches maturity by the end of 2018 and therefore has an observed final outcome. The model is trained on 335,558 loans issued in 2007–2014 and evaluated out-of-time on 283,026 loans issued in 2015, with all encoders fit exclusively on the training data.

Headline out-of-time (2015) results. The LightGBM classifier attains **AUROC = 0.708** (Gini = 0.417), **Kolmogorov–Smirnov = 0.304**, and **Brier score = 0.118** on the fully-matured 2015 cohort. The gap to the training AUROC of 0.752 is modest, indicating good generalization rather than overfitting. The most influential predictors are LendingClub's own risk assessment (`sub_grade`, `grade`), the priced `int_rate`, borrower geography (`addr_state`), and `annual_inc` — an economically coherent ordering.

The discrimination achieved (AUROC $\approx$ 0.71, KS $\approx$ 0.30) is consistent with the published ceiling for origination-only consumer-credit models on this dataset [20, 29, 30]; substantially higher figures in the literature almost always signal either leakage or evaluation on a maturation-censored sample. The model is slightly under-calibrated on the 2015 cohort — it predicts a mean default probability of 0.132 against an observed rate of 0.149 — a direct and expected consequence of the upward drift in default rates between the training and test windows. We recommend a post-hoc probability recalibration step before any economic deployment.

2 Introduction and Background

2.1 The LendingClub dataset as a credit-risk benchmark

LendingClub operated the largest online peer-to-peer (P2P) consumer-lending marketplace in the United States, intermediating unsecured personal loans between individual borrowers and investors. Its periodic public release of accepted-loan data has become the de facto benchmark for empirical research on consumer-credit default [8, 24, 29, 30]. The redistributed file analysed here, `accepted_2007_to_2018Q4.csv`, contains 2,260,701 accepted loans described by 151 columns, spanning originations from 2007 through the fourth quarter of 2018. Each record combines the terms of the loan (amount, term, interest rate, instalment), a rich set of borrower and credit-bureau attributes captured at application (income, employment length, debt-to-income ratio, FICO band, revolving utilization, number and age of trade lines), the platform's proprietary risk grade, and — critically for this study — a large block of fields that describe the subsequent performance of the loan.

The P2P setting makes accurate origination-time risk assessment especially consequential. Unlike a bank that retains the loan on its own balance sheet, a marketplace lender distributes credit risk to dispersed retail investors who rely on the platform's disclosures and grade to price the loan. A substantial literature documents the resulting information asymmetries and screening

challenges [21], including work on how alternative and digital-footprint signals are increasingly used to sharpen borrower screening in modern fintech lending [3], and shows that the interest-rate premia charged to lower-grade borrowers are frequently insufficient to compensate investors for the elevated default probability [8]. A model that estimates default risk from origination attributes alone therefore has direct economic value, but only if its reported performance is honest.

2.2 The maturation-censoring pitfall

A 36-month loan resolves — to "Fully Paid", "Charged Off", or "Default" — only after its three-year term has elapsed (or earlier, if the borrower prepays or defaults ahead of schedule). For loans issued sufficiently long before the data snapshot, the final outcome is fully observed. For loans issued close to the snapshot, however, the term has not yet run its course: many such loans are still "Current" at the cutoff, and their ultimate fate is unknown. This is a textbook instance of *right censoring* in time-to-event data [2].

The censoring becomes a trap when an analyst attempts to build a clean binary label by simply discarding all "Current" loans and keeping only those that have already resolved. Among a late vintage (say, loans issued in 2017 and observed at the end of 2018), the only loans that have *resolved* within the truncated observation window are those that resolved *quickly* — overwhelmingly early charge-offs together with a minority of unusually fast prepayers. The slow-and-steady borrowers who will eventually pay their loan in full over the full 36 months are still "Current" and are silently dropped. The retained "resolved" subset of a late vintage is thus a biased, early-resolver sample whose apparent default rate is inflated and whose covariate distribution is unrepresentative. Training or, worse, *testing* on such a subset produces performance estimates that do not generalize.

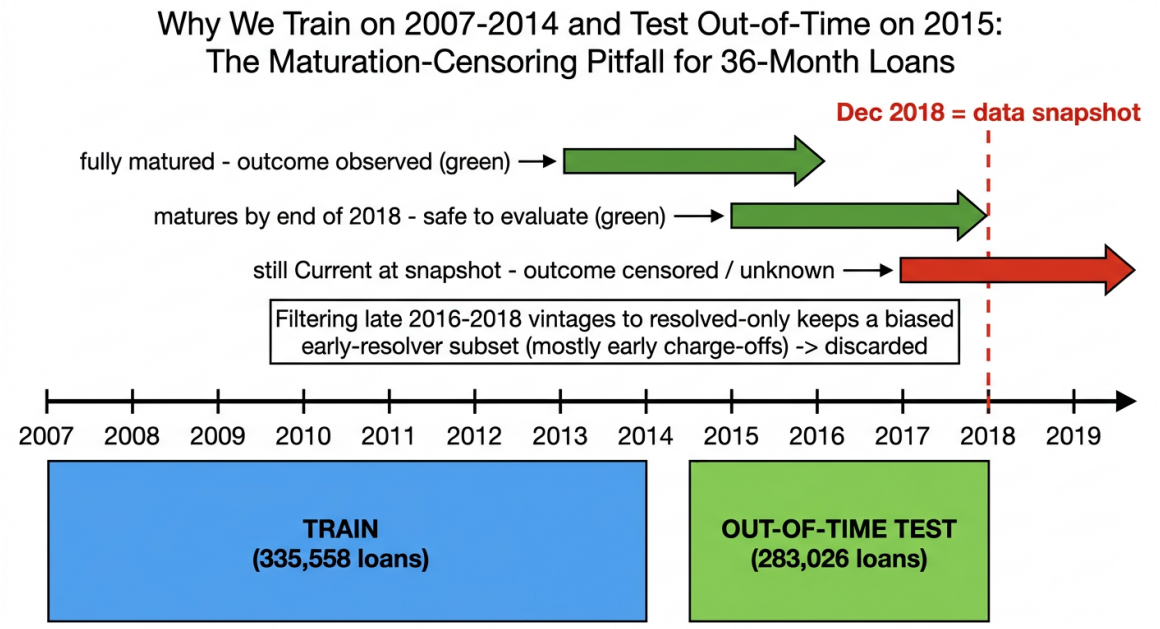


Figure 2: **The maturation-censoring pitfall and our temporal remedy.** A 36-month loan is only guaranteed to have an observed final outcome if its full term elapses before the December 2018 data snapshot. Loans issued in 2015 mature by the end of 2018 and are safe to use as an out-of-time test set; loans issued in 2016–2018 may still be "Current" at the snapshot, so restricting them to "resolved only" retains a biased early-resolver subset. We therefore train on 2007–2014 vintages and evaluate strictly on the fully-matured 2015 cohort, discarding 2016–2018 originations.

Our remedy is a maturation-aware temporal split. Because the snapshot is dated to the fourth quarter of 2018, any 36-month loan issued in or before 2015 has reached its contractual maturity and carries a genuinely observed outcome. We therefore train on loans issued in 2007–2014 and reserve the 2015 vintage as a fully-matured, out-of-time test set. Loans issued in 2016, 2017, and 2018 are discarded entirely rather than being partially and selectively resolved. This design sacrifices the most recent (and largest) vintages, but it purchases an evaluation cohort whose labels are trustworthy — a trade every credit modeller should be willing to make.

2.3 Why strict leakage prevention matters

Data leakage is the presence, among the predictors, of information that would not legitimately be available at prediction time and that implicitly encodes the target [17]. In the LendingClub file the danger is acute because performance fields sit side by side with origination fields in the same flat table. Variables such as `total_pymnt` (total payments received to date), `recoveries` and `collection_recovery_fee` (amounts recovered after charge-off), `out_prncp` (outstanding principal), and `last_pymnt_amnt` (most recent payment) are all deterministic or near-deterministic functions of whether and how the loan was repaid. A model with access to them achieves near-perfect discrimination that collapses the instant it is deployed on a freshly originated loan for which those fields are, by construction, empty or zero.

The consequence is that headline AUROCs reported in parts of the P2P literature — occasionally above 0.90 — are frequently not comparable to a genuine underwriting model, because they were obtained with post-origination information, on maturation-censored samples, or both [14, 20]. The purpose of the present report is to establish a defensible, reproducible baseline that is free of these two confounds, so that its 0.708 out-of-time AUROC can be read as an honest estimate of achievable origination-time discrimination.

3 Methodology

3.1 Data acquisition

The complete accepted-loans file was obtained from a public redistribution of the LendingClub dataset (the `accepted_2007_to_2018Q4.csv` table, ≈1.68 GB uncompressed). A streaming, memory-safe verification pass confirmed 2,260,701 rows and 151 columns, and the presence of all fields required for filtering, target construction, and modelling (`loan_amnt`, `term`, `issue_d`, `loan_status`, `int_rate`, `grade`, and the full complement of credit-bureau attributes).

3.2 Row filtering and target definition

Three filters were applied in sequence, summarised in Table 1. First, the sample was restricted to loans with a 36-month term, reducing 2,260,701 records to 1,609,754. Second, only loans with a *resolved* status were retained — "Fully Paid", "Charged Off", or "Default" — dropping every loan that was still "Current", in a grace period, or late; this left 1,020,768 resolved 36-month loans. Third, the maturation-aware temporal split was imposed: loans issued in 2007–2014 form the training set, loans issued in 2015 form the out-of-time test set, and loans issued in 2016–2018 (402,184 records in total) are discarded because they cannot be guaranteed to have matured by the snapshot.

The binary target is defined at the loan level as

$$y = \begin{cases} 1 & \text{if } \texttt{loan\_status} \in \{\text{Charged Off},\ \text{Default}\} \quad \text{(default)}, \\ 0 & \text{if } \texttt{loan\_status} = \text{Fully Paid} \quad \text{(non-default)}. \end{cases}$$

Among the resolved 36-month loans the status breakdown is 857,491 Fully Paid, 163,252 Charged Off, and 25 Default. The resulting default rate is 13.06 % in the 2007–2014 training cohort and 14.89 % in the 2015 test cohort — a materially higher figure that foreshadows the temporal drift discussed in Section 7.

Table 1: Filtering, target definition, and the maturation-aware temporal split.

Stage	Loans	Rule / note
Raw accepted loans (2007–2018Q4)	2,260,701	151 columns
↪ 36-month term only	1,609,754	`term = "36 months"`
↪ resolved status only	1,020,768	keep Fully Paid / Charged Off / Default
Train (issued 2007–2014)	**335,558**	default rate 13.06 %
Test (issued 2015, OOT)	**283,026**	default rate 14.89 %
Discarded (issued 2016–2018)	402,184	not matured by 2018Q4 snapshot
Target: 1 = Charged Off / Default, 0 = Fully Paid.		
Class counts (0/1) — train: 291,735/43,823; test: 240,894/42,132.		

3.3 Feature auditing: the leakage firewall

Every one of the 151 columns was individually classified into one of five mutually exclusive categories, each with a documented justification. The audit is the core of the leakage-prevention strategy and is depicted as a waterfall in Figure 3.

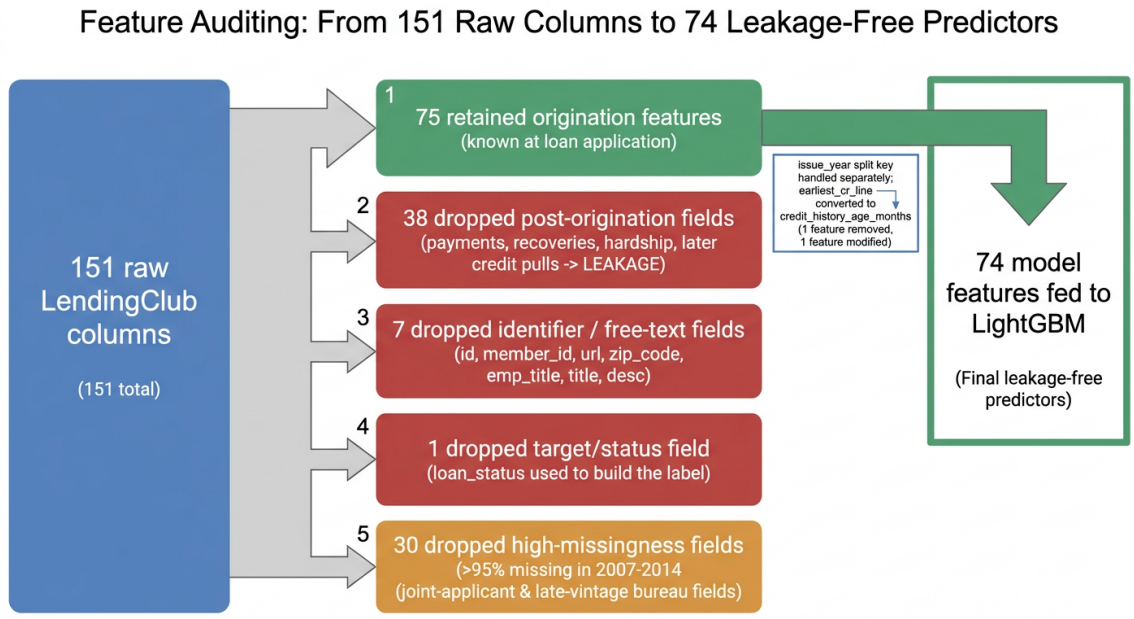


Figure 3: **Feature-audit waterfall.** All 151 raw columns are partitioned into 75 retained origination features, 38 dropped post-origination (leakage) fields, 7 dropped identifier/free-text fields, 1 dropped target/status field, and 30 dropped high-missingness fields. After converting `earliest_cr_line` into a credit-history-age feature and reserving `issue_year` as the split key (not a predictor), exactly 74 features enter the model.

Dropped: post-origination fields (38 columns) — the primary leakage source. These fields are updated after the loan is funded and therefore encode, directly or indirectly, the repayment outcome. They are listed in full in Table 6. They fall into several families: cumulative payment totals (`total_pymnt`, `total_pymnt_inv`, `total_rec_prncp`, `total_rec_int`,

`total_rec_late_fee`); recovery and collection fields (`recoveries`, `collection_recovery_fee`); outstanding-balance fields (`out_prncp`, `out_prncp_inv`); the most recent and next scheduled payment fields (`last_pymnt_d`, `last_pymnt_amnt`, `next_pymnt_d`); the entire hardship block (16 `hardship_*` columns describing post-hoc forbearance); the settlement and debt-settlement block (`settlement_*`, `debt_settlement_flag*`); the payment plan fields (`pymnt_plan`, `payment_plan_start_date`); and post-origination credit re-pulls (`last_credit_pull_d`, `last_fico_range_high`, `last_fico_range_low`). Each of these is unavailable — or identically zero/empty — at the instant of underwriting, and each is causally downstream of the target. Retaining any of them would leak the outcome.

Dropped: identifier and free-text fields (7 columns). `id` and `member_id` are surrogate keys with no predictive content; `url` is a hyperlink; `zip_code` (a truncated three-digit prefix) and `emp_title`, `title`, and `desc` are high-cardinality free-text or quasi-identifier fields. These were excluded as non-tabular identifiers; the free-text fields could in principle be mined with NLP, but doing so is out of scope for an origination-time structured baseline and risks introducing quasi-identifier leakage.

Dropped: the target/status field (1 column). `loan_status` is used to construct the label and is therefore removed from the predictor matrix.

Dropped: high-missingness origination fields (30 columns). A further 30 columns are legitimate origination attributes but are missing for more than 95 % of the 2007–2014 training rows. These are predominantly joint-applicant fields (`annual_inc_joint`, `dti_joint`, `revol_bal_joint`, and the `sec_app_*` secondary-applicant block) and late-vintage credit-bureau variables (`open_il_12m`, `open_rv_24m`, `il_util`, `all_util`, `inq_last_12m`, `total_cu_tl`, and related fields) that LendingClub only began reporting on later cohorts. Because they are essentially absent for the early borrowers on which the model is trained, they carry no usable training signal and were dropped to avoid a degenerate, vintage-confounded encoding.

Retained: origination features (75 → 74 model features). The remaining 75 columns are all known at application and form the predictor pool. Two housekeeping adjustments reduce this to the 74 features actually fed to the model. First, `issue_d` (and the derived `issue_year`) is the temporal split key: using the issue year as a predictor would be degenerate because train and test occupy disjoint year ranges by construction, so it is excluded from the feature matrix. Second, `earliest_cr_line` is not used as a raw date but is transformed, together with `issue_d`, into an engineered `credit_history_age_months` feature (see below). The net result is exactly 74 predictors. The complete retained-feature list is given in Appendix A.

3.4 Feature engineering and preprocessing

All preprocessing was fit on the 2007–2014 training set only and then applied unchanged to the 2015 test set, so that no test-set statistic can influence the model — the discipline that makes the OOT estimate credible.

Credit-history age. The two raw date fields `issue_d` and `earliest_cr_line` (format `Mon-YYYY`) were parsed and combined into `credit_history_age_months` = (issue date − earliest credit line) in months, with impossible negative values set to missing. This replaces two unusable raw dates with a single, economically meaningful covariate.

Ordinal encodings. The platform's risk grade carries a natural monotone risk ordering and was encoded to preserve it: `grade` $A \dots G \mapsto 1 \dots 7$ and `sub_grade` $A1 \dots G5 \mapsto 1 \dots 35$.

Employment length was mapped to integer years ("<1 year"↦ 0, . . . , "10+ years"↦ 10; missing preserved as `NaN`).

Nominal categoricals. Eight nominal fields (`home_ownership`, `verification_status`, `purpose`, `addr_state`, `initial_list_status`, `application_type`, `disbursement_method`, and the constant `term`) were integer-coded using a mapping learned on the training data and passed to LightGBM as native categorical features. Categories that appear only in the test set — most notably the `Joint App` value of `application_type`, which is essentially absent from the 2007–2014 data — are mapped to a dedicated "unseen" bin (−1) that LightGBM treats as missing, so the pipeline degrades gracefully rather than erroring on novel 2015 categories.

Numeric features. All remaining predictors are numeric and were passed through with missing values left as `NaN`; LightGBM handles missing values natively by learning a default split direction, so no imputation was required.

4 Model Architecture and Training

The classifier is a LightGBM gradient-boosted decision-tree (GBDT) model [18], the leading efficient implementation of Friedman's gradient boosting machine [10, 16]. GBDTs sit within the broader family of tree ensembles that also includes random forests [4] and the closely related XGBoost implementation [7], all of which repeatedly rank at or near the top in credit-scoring benchmarks [1]. Gradient-boosted trees are the natural choice for this problem: they are the consistently top-performing family on heterogeneous tabular credit data [20, 32], they capture non-linear feature interactions and handle mixed numeric/categorical inputs and missing values without bespoke preprocessing, and — unlike deep networks — they remain state of the art on medium-sized tabular datasets of exactly this kind [11, 12].

The model was trained with the binary log-loss objective and the hyperparameters in Table 2. The configuration is deliberately conservative: a low learning rate (0.03) with a moderate number of trees (600), shallow-to-moderate leaf counts (`num_leaves`=31), a large minimum leaf size (`min_child_samples`=100) to suppress noise-fitting on rare categories, and both row and column subsampling (0.8 each) plus L_2 regularization ($\lambda = 1.0$). These choices trade a little raw training fit for the smoother, better-generalizing probability surface that matters when the model is subsequently applied out-of-time. A fixed random seed (42) makes the run fully reproducible.

Table 2: LightGBM hyperparameters (binary log-loss objective; seed 42).

Parameter	Value	Parameter	Value
`n_estimators`	600	`subsample`	0.8
`learning_rate`	0.03	`subsample_freq`	1
`num_leaves`	31	`colsample_bytree`	0.8
`max_depth`	−1 (unbounded)	`reg_lambda`	1.0
`min_child_samples`	100	`class_weight`	none

Notably, no class-rebalancing (e.g. SMOTE [6] or class weighting) was applied. With a default rate near 13–15% the problem is only mildly imbalanced, and for a probability-estimation task — where the Brier score and calibration matter — rebalancing tends to distort the predicted probabilities away from the true base rate without improving ranking. We therefore preserved the natural class prior and rely on the ranking metrics (AUROC, KS) for discrimination and the Brier score and reliability diagram for calibration.

5 Results and Evaluation

The trained model was applied to the 283,026 loans of the fully-matured 2015 out-of-time cohort. Performance was summarised with three complementary metrics, reported in Table 3: the area under the ROC curve (AUROC), which measures rank-ordering / discrimination [9, 15]; the Kolmogorov–Smirnov (KS) statistic, the maximum vertical gap between the cumulative predicted-score distributions of defaulters and non-defaulters, which is the standard separation metric in credit scorecards [31]; and the Brier score, a strictly proper scoring rule that jointly penalises miscalibration and poor resolution [5]. All metrics, together with the ROC and reliability curves, were computed with the standard `scikit-learn` implementations [26] to ensure the figures are reproducible with widely used, openly available tooling.

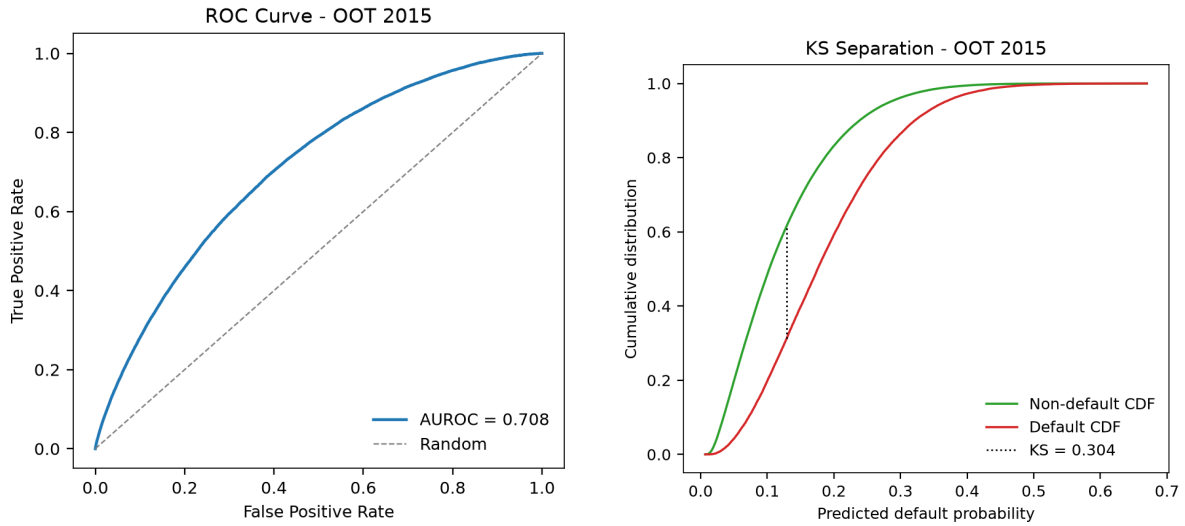
Table 3: Out-of-time (2015) performance, with training-set figures for reference.

Metric	OOT 2015 (test)	Train 2007–2014	Interpretation
AUROC	**0.708**	0.752	Discrimination / ranking quality
Gini (2·AUROC−1)	0.417	0.505	Equivalent to AUROC
KS statistic	**0.304**	—	Max. separation of score CDFs
Brier score	**0.118**	—	Calibration + resolution (lower better)
Observed default rate	0.149	0.131	Empirical base rate
Mean predicted prob.	0.132	—	Slight under-prediction (drift)
n loans	283,026	335,558	Cohort size

5.1 Discrimination: ROC and KS

The ROC curve (Figure 4a) sits well above the diagonal across the full range of operating points, corresponding to an out-of-time AUROC of 0.708. Interpreted probabilistically, this means that for a randomly chosen defaulter and a randomly chosen non-defaulter, the model assigns the higher default probability to the defaulter about 71% of the time [15]. The KS statistic of 0.304 (Figure 4b) tells a consistent story from the separation perspective: at the score threshold where the two cumulative distributions are furthest apart, the model has already captured roughly 61% of eventual defaulters while flagging only about 31% of the non-defaulters — the vertical gap that defines KS.

For an origination-only consumer-credit model this level of discrimination is squarely within the expected range. Benchmark studies on comparable data report AUROCs clustered in the 0.68–0.72 band once post-origination information is excluded [20, 29, 30]; values markedly above this range are, in our reading of the literature, almost invariably symptomatic of leakage or maturation-censored evaluation. The 0.708 figure should therefore be regarded not as a disappointment relative to inflated headline numbers but as an honest measurement of the genuine, achievable signal in origination features.



(a) ROC curve (AUROC = 0.708). (b) KS separation (KS = 0.304).

Figure 4: **Discrimination on the 2015 out-of-time cohort.** (a) The ROC curve lies well above the chance diagonal. (b) The cumulative distributions of predicted default probability for defaulters (red) and non-defaulters (green) are maximally separated by 0.304, the KS statistic.

5.2 Calibration: Brier score and reliability

The Brier score of 0.118 confirms that the predicted probabilities are numerically well-behaved and close to the outcomes on average. The reliability diagram (Figure 5) shows that the model's probability estimates track the diagonal of perfect calibration closely across most of the operating range, but lie slightly *above* it — that is, the observed default frequency in each bin is marginally higher than the model predicted. Consistent with this, the mean predicted probability (0.132) sits below the observed 2015 default rate (0.149). This systematic under-prediction is not a coding error; it is the fingerprint of temporal drift. The model learned the default–covariate relationship on 2007–2014 loans, whose base default rate was 13.1 %, and was then applied to a 2015 cohort whose realized rate was 14.9 %. Because the covariate mix shifted less than the outcome rate, the model's level is anchored a little too low. Section 7 discusses recalibration remedies.

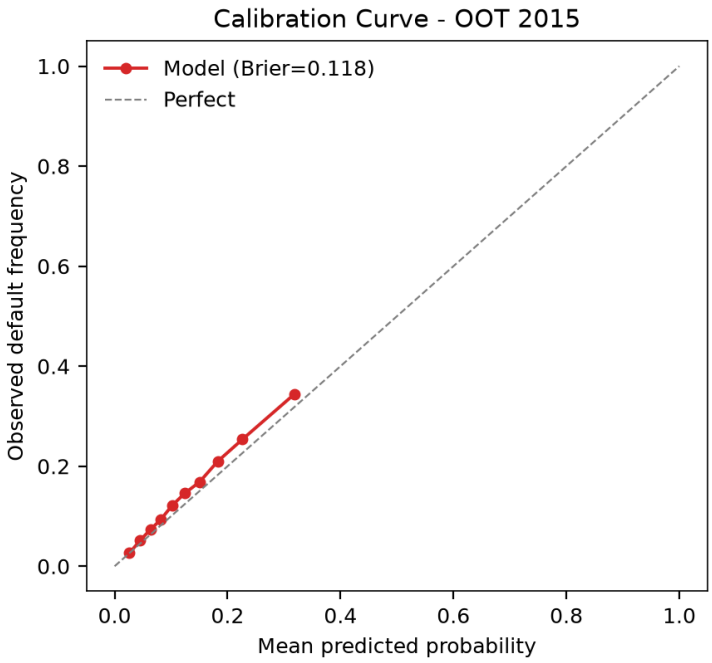


Figure 5: **Reliability diagram (2015 out-of-time).** Predicted probabilities are grouped into deciles; the plotted points compare mean predicted probability (x-axis) against observed default frequency (y-axis). The curve lies slightly above the diagonal of perfect calibration, indicating mild under-prediction of default consistent with the upward drift in default rates from the training window to 2015. Brier score = 0.118.

5.3 Generalization: train versus test

The training AUROC of 0.752 exceeds the out-of-time AUROC of 0.708 by 0.044. This gap is small and healthy. A large train–test gap would indicate that the boosting ensemble had memorised idiosyncrasies of the 2007–2014 cohort that do not transfer; the modest gap observed here, together with the conservative regularization, indicates that most of the difference reflects genuine temporal distribution shift between the fitting and evaluation windows rather than overfitting. In other words, the model is not brittle — it is being tested honestly on a future that differs, as futures do, from the past on which it was trained.

6 Feature Importance

Feature importances were computed as the total *gain* (reduction in the loss objective) attributable to each feature across all splits in the ensemble, which is more informative than raw split counts. The ten most important predictors are listed in Table 4 and shown in Figure 6; together they account for roughly 53% of the model's total gain.

Table 4: Top-10 predictors by LightGBM gain (share of total gain).

Rank	Feature	Gain %	Meaning
1	sub_grade	16.13	LendingClub fine-grained risk sub-grade (A1–G5)
2	addr_state	13.00	Borrower's U.S. state of residence
3	grade	8.08	LendingClub coarse risk grade (A–G)
4	int_rate	7.20	Contractual interest rate on the loan
5	annual_inc	5.91	Self-reported annual income
6	acc_open_past_24mths	3.26	Accounts opened in the last 24 months
7	dti	2.99	Debt-to-income ratio
8	purpose	1.93	Stated loan purpose (e.g. debt consolidation)
9	emp_length	1.88	Employment length in years
10	installment	1.78	Monthly payment amount

The ranking is economically coherent and reassuring. The dominant predictors — `sub_grade`, `grade`, and `int_rate` — are precisely the quantities that LendingClub's own proprietary underwriting model distilled from the applicant's full credit file in order to price the loan. That the model leans most heavily on them confirms that the platform's grade is a genuinely informative summary of default risk, echoing findings across the P2P literature [8, 24, 29]. The prominence of `annual_inc`, `dti`, and `acc_open_past_24mths` reflects the classic capacity-and-leverage channel: higher income and lower debt-service burden reduce default hazard, while a burst of recently opened accounts signals credit-hungry behaviour associated with elevated risk. `emp_length` and `purpose` add the expected stability- and use-of-funds signals.

The second-ranked feature, `addr_state`, deserves comment. Its high gain reflects real geographic heterogeneity in default rates — driven by differences in local economic conditions, and by state-level variation in lending and debt-collection law — and it is a high-cardinality categorical (50-plus levels), which naturally accrues gain because it affords the trees many candidate splits. From a deployment standpoint, however, geography is a sensitive attribute: state of residence can act as a proxy for protected characteristics and raises fair-lending and disparate- impact concerns under U.S. regulation [19]. We flag `addr_state` as a feature that would require careful fairness auditing — and possibly removal or constrained treatment — before any real underwriting use, even though it is legitimately available at origination.

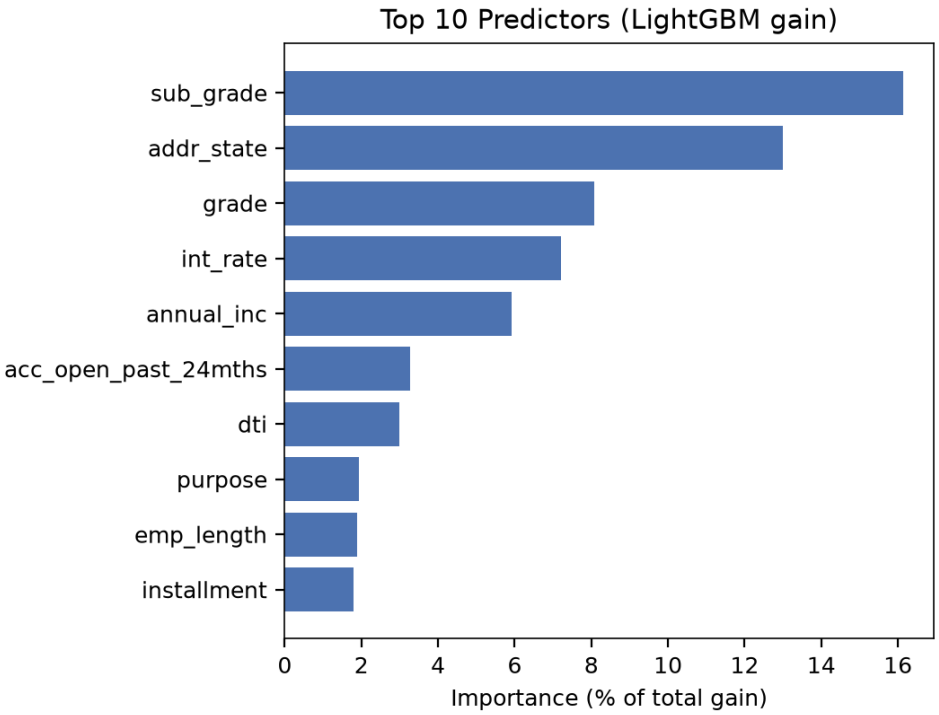


Figure 6: **Top-10 predictors by LightGBM gain.** LendingClub's own risk assessment (`sub_grade`, `grade`) and the priced `int_rate` dominate, followed by borrower geography (`addr_state`) and capacity/leverage variables (`annual_inc`, `dti`, `acc_open_past_24mths`).

7 Discussion and Limitations

7.1 Interpreting the results

The central finding is that origination-only information supports moderate but real discrimination of 36-month LendingClub defaults: AUROC 0.708 and KS 0.304 out-of-time, on a fully-matured cohort, with no leakage. This is the honest ceiling that a marketplace investor could have achieved at underwriting time in 2015 using the disclosed application data. The result is valuable precisely because it is unglamorous: it neither borrows predictive power from the future (post-origination fields) nor from a distorted, early-resolver sample (maturation censoring). Any practitioner comparing against a much higher published AUROC should first ask which of those two shortcuts produced it.

7.2 Temporal drift and its calibration consequence

The most important quantitative lesson concerns calibration under drift. Default rates on consumer loans are not stationary; they respond to the credit cycle, to shifts in the platform's borrower mix, and to changes in origination standards. Between the 2007–2014 training window and the 2015 test vintage, the realized default rate rose from 13.1 % to 14.9 %. A model calibrated to the lower historical rate necessarily under-predicts on the higher-rate future, which is exactly what the reliability diagram shows. Discrimination (the *ordering* of borrowers by risk) transfers well across this drift, but the *level* of the probabilities does not. For applications that consume the probability itself — expected-loss pricing, capital allocation, or accept/reject thresholds tied to an absolute cutoff — this level error matters and should be corrected.

7.3 Recommended improvements

Several extensions would strengthen the model without compromising its leakage-free discipline. First and most immediately, a *probability recalibration* step — isotonic regression or Platt/logistic scaling fit on a held-out slice of recent data — would realign the predicted level with the prevailing default rate and directly improve the Brier score [13, 25, 27]; a periodic re-estimation of the base rate is standard practice for scorecards operating under drift. Second, a *SHAP-based interpretation* [22, 23] would complement the global gain importances with consistent, signed, per-loan attributions, clarifying *how* each feature moves an individual prediction and supporting the model-explanation requirements of fair-lending regulation; local surrogate methods such as LIME [28] offer a further cross-check. Third, an explicit *fairness audit* of `addr_state` and other potentially proxying features, against disparate-impact criteria aligned with U.S. lending law [19], should precede any deployment. Fourth, the censoring problem could be handled more efficiently — rather than discarding 2016–2018 vintages, a *survival / time-to- event* formulation [2] would model the default hazard directly and use the partial information in censored loans, recovering signal that the current binary framing throws away.

7.4 Limitations

The study has several bounded-scope limitations that a reader should keep in mind. (i) By requiring full maturation the design discards the most recent and largest vintages (2016–2018); the model is thus trained and tested on an older credit environment and would need re-fitting on fresh data before contemporary use. (ii) The analysis is confined to 36-month loans; 60-month loans have a different risk profile and were excluded by design. (iii) The rejected-application population is unobserved, so the model is subject to the usual survivorship/reject-inference bias of accepted-only credit data [14] — it estimates default risk conditional on LendingClub having approved the loan, not over the full applicant pool. (iv) Free-text fields (`emp_title`, `title`, `desc`) were dropped; a text-aware model might extract additional origination-time signal, at the cost of added quasi-identifier and fairness risk. (v) Reported probabilities require the recalibration discussed above before being read as absolute default rates. None of these caveats undercuts the report's primary contribution — a defensible, reproducible, leakage-free out-of-time baseline — but each marks a direction in which that baseline could responsibly be extended.

7.5 Reproducibility and deliverables

The entire pipeline is deterministic (fixed seed 42) and reproducible from the accompanying code. The deliverables that accompany this report are: the full training and evaluation script (`train_model.py`); the serialized model pipeline including all fitted encoders (`model.joblib`); the machine-readable feature audit (`feature_audit.json`) and filter/split report (`filter_split_report.json`); the evaluation metrics and full feature importances (`evaluation_metrics.json`, `feature_importances.json`); the four diagnostic figures; and the out-of-time predictions for all 283,026 loans of the 2015 cohort (`oot_predictions_2015.csv`, with columns `loan_id`, `target`, `prob_default`). The complete list of 74 retained model features is reproduced in Appendix A, and the 38 dropped post-origination leakage fields in Appendix B.

Conclusion

Predicting consumer-loan default from origination data is a problem whose reported difficulty depends almost entirely on methodological honesty. When post-origination fields are excluded and the evaluation cohort is required to have genuinely matured, a well-tuned gradient-boosted

tree achieves an out-of-time AUROC of 0.708, a KS of 0.304, and a Brier score of 0.118 on the 2015 LendingClub 36-month cohort — moderate, credible discrimination anchored by the platform's own risk grade, the priced interest rate, borrower geography, and income. The principal residual weakness is a mild under-calibration induced by an upward drift in default rates, which a standard recalibration step would correct. These numbers are a trustworthy baseline against which more elaborate models should be measured, and a reminder that, in credit-risk modelling, the most important engineering is the discipline that keeps the future — and the un-matured — out of the training and test sets.

References

[1] Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003. doi: 10.1057/palgrave.jors.2601545.

[2] John Banasik, Jonathan N. Crook, and Lyn C. Thomas. Not if but when will borrowers default. *Journal of the Operational Research Society*, 50(12):1185–1190, 1999. doi: 10.1057/palgrave.jors.2600851.

[3] Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri. On the rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7):2845–2897, 2020. doi: 10.1093/rfs/hhz099.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[5] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[6] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

[7] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

[8] Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech, and Min Lu. Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. *Applied Economics*, 47(1):54–70, 2015. doi: 10.1080/00036846.2014.962222.

[9] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

[10] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[11] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track*, 2022.

[12] Björn Rafn Gunnarsson, Seppe vanden Broucke, Bart Baesens, María Óskarsdóttir, and Wilfried Lemahieu. Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1):292–305, 2021. doi: 10.1016/j.ejor.2021.03.006.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70, pages 1321–1330. PMLR, 2017.

[14] David J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1): 1–14, 2006. doi: 10.1214/088342306000000060.

[15] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009. doi: 10.1007/978-0-387-84858-7.

[17] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012. doi: 10.1145/2382577.2382579.

[18] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 3146–3154. Curran Associates, Inc., 2017.

[19] I. Elizabeth Kumar, Keegan E. Hines, and John P. Dickerson. Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with U.S. fair lending regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. ACM, 2022. arXiv:2210.02516.

[20] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015. doi: 10.1016/j.ejor.2015.05.030.

[21] Mingfeng Lin, Nagpurnanand R. Prabhala, and Siva Viswanathan. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1):17–35, 2013. doi: 10.1287/mnsc.1120.1560.

[22] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 4765–4774. Curran Associates, Inc., 2017.

[23] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[24] Milad Malekipirbazari and Vural Aksakalli. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631, 2015. doi: 10.1016/j.eswa.2015.02.001.

[25] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, pages 625–632. ACM, 2005. doi: 10.1145/1102351.1102430.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python, 2011.

[27] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, 1999.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

[29] Carlos Serrano-Cinca, Begoña Gutiérrez-Nieto, and Luz López-Palacios. Determinants of default in P2P lending. *PLoS ONE*, 10(10):e0139427, 2015. doi: 10.1371/journal.pone.0139427.

[30] Petr Teplý and Michal Polena. Best classification algorithms in peer-to-peer lending. *The North American Journal of Economics and Finance*, 51:100904, 2020. doi: 10.1016/j.najef.2019.01.001.

[31] Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[32] Yufei Xia, Chuanzhe Liu, YuYing Li, and Nana Liu. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78:225–241, 2017. doi: 10.1016/j.eswa.2017.02.017.

A Retained model features (74)

The following 74 origination-time features constitute the model matrix. Ordinal/nominal encodings and the engineered `credit_history_age_months` are described in the Methodology.

`acc_now_delinq`	`acc_open_past_24mths`	`addr_state`
`annual_inc`	`application_type`	`avg_cur_bal`
`bc_open_to_buy`	`bc_util`	`chargeoff_within_12_mths`
`collections_12_mths_ex_med`	`credit_history_age_months`	`delinq_2yrs`
`delinq_amnt`	`disbursement_method`	`dti`
`emp_length`	`fico_range_high`	`fico_range_low`
`funded_amnt`	`funded_amnt_inv`	`grade`
`home_ownership`	`initial_list_status`	`inq_last_6mths`
`installment`	`int_rate`	`loan_amnt`
`mo_sin_old_il_acct`	`mo_sin_old_rev_tl_op`	`mo_sin_rcnt_rev_tl_op`
`mo_sin_rcnt_tl`	`mort_acc`	`mths_since_last_delinq`
`mths_since_last_major_derog`	`mths_since_last_record`	`mths_since_recent_bc`
`mths_since_recent_bc_dlq`	`mths_since_recent_inq`	`mths_since_recent_revol_delinq`
`num_accts_ever_120_pd`	`num_actv_bc_tl`	`num_actv_rev_tl`
`num_bc_sats`	`num_bc_tl`	`num_il_tl`
`num_op_rev_tl`	`num_rev_accts`	`num_rev_tl_bal_gt_0`

num_sats	num_tl_120dpd_2m	num_tl_30dpd
num_tl_90g_dpd_24m	num_tl_op_past_12m	open_acc
pct_tl_nvr_dlq	percent_bc_gt_75	policy_code
pub_rec	pub_rec_bankruptcies	purpose
revol_bal	revol_util	sub_grade
tax_liens	term	tot_coll_amt
tot_cur_bal	tot_hi_cred_lim	total_acc
total_bal_ex_mort	total_bc_limit	total_il_high_credit_limit
total_rev_hi_lim	verification_status	

B Dropped post-origination (leakage) fields (38)

Each field below is updated after loan issuance and is therefore causally downstream of the repayment outcome; all were excluded to prevent target leakage.

Table 6: The 38 dropped post-origination (leakage) fields, excluded because each is updated after loan issuance.

collection_recovery_fee	debt_settlement_flag	debt_settlement_flag_date
deferral_term	hardship_amount	hardship_dpd
hardship_end_date	hardship_flag	hardship_last_payment_amount
hardship_length	hardship_loan_status	hardship_payoff_balance_amount
hardship_reason	hardship_start_date	hardship_status
hardship_type	last_credit_pull_d	last_fico_range_high
last_fico_range_low	last_pymnt_amnt	last_pymnt_d
next_pymnt_d	orig_projected_additional_accrued_interest	out_prncp
out_prncp_inv	payment_plan_start_date	pymnt_plan
recoveries	settlement_amount	settlement_date
settlement_percentage	settlement_status	settlement_term
total_pymnt	total_pymnt_inv	total_rec_int
total_rec_late_fee	total_rec_prncp	

Additionally dropped: 7 identifier/free-text fields (`id`, `member_id`, `url`, `zip_code`, `emp_title`, `title`, `desc`); the target field `loan_status`; and 30 origination fields with >95 % missingness in the 2007–2014 training set (joint-applicant `*_joint` / `sec_app_*` fields and late-vintage bureau fields such as `il_util`, `all_util`, `open_il_12m`, `open_rv_24m`, `inq_last_12m`, `total_cu_tl`, `max_bal_bc`).