International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# Risk Decisioning for Unsecured Lending Using Gradient Boosted Decision Trees: Calibration, Governance, and Operational Trade-Offs

**Ajay Punia**

**Abstract:** ***Unsecured lending requires credit decisions under asymmetric error costs and strict governance constraints. Gradient boosted decision trees (GBDTs) are increasingly used for default-risk estimation because they capture nonlinearities, interaction effects, and heterogeneous feature types better than classical scorecards in many settings. This paper presents a decisioning-focused research blueprint for using GBDTs in unsecured lending, covering data definition, leakage control, temporal validation, probability calibration, threshold design, and explanation practices suitable for regulated credit workflows. The approach combines a GBDT probability-of-default model with a separate calibration layer and a policy engine that maps calibrated risk to approve/refer/decline outcomes and, where applicable, risk-based pricing or limit assignment. Hypothetical but operationally plausible results suggest improved rank-ordering performance and better tail separation than a logistic baseline, while highlighting that raw boosted outputs can be miscalibrated and require post-hoc correction. The discussion emphasizes stability under portfolio drift, reject-inference bias, and the practical limits of explainability and fairness criteria in credit. The paper concludes with research directions on robust learning under selection bias, constrained boosting, and drift-aware calibration.***

**Keywords:** credit risk, unsecured lending, gradient boosted decision trees, calibration, explainability

## 1. Introduction

Unsecured consumer lending is a high-frequency decision environment where lenders must assess repayment risk without collateral. The decision is not simply "predict default," but "allocate credit under uncertainty," balancing expected profitability, customer value, and regulatory expectations. In practice, approval policies and pricing structures depend on a model's estimated probability of default (PD), the expected severity of loss when default occurs, and operational expenses linked to verification, servicing, and collections. Because these components behave differently across borrower segments and macroeconomic regimes, the decision system must be accurate, stable, and understandable.

Historically, logistic regression scorecards have served as the default modeling choice due to their transparency and governance convenience (Hand & Henley, 1997; Thomas, 2000; Thomas et al., 2002). Yet, decades of comparative work show that nonlinear learners can provide meaningful gains in rank-ordering performance when feature sets are rich or interactions are strong (Baesens et al., 2003; Lessmann et al., 2015). The practical motivation is clear: better separation in the risk tails enables tighter policy controls, more selective verification, and less noisy adverse selection.

Gradient boosted decision trees (GBDTs) occupy a particularly useful point on the modeling spectrum. Compared with single trees, boosting reduces bias and improves predictive strength through sequential fitting (Friedman, 2001). GBDTs are usually easier to train and use for tabular credit features than deep neural networks. With careful regularization, they can also give good results (Chen & Guestrin, 2016; Ke et al., 2017). Simultaneously, boosting presents significant challenges in lending: initial calibration of probability estimates may be inadequate (Niculescu-Mizil & Caruana, 2005), decision policies may become sensitive to slight changes in distribution, and explanation outputs may not align well with compliant reason codes (Ribeiro et al., 2016; Lundberg & Lee, 2017).

This paper frames "Using gradient-boosted decision trees-based risk decisions in unsecured lending" as an end-to-end decision problem rather than an isolated classification exercise. The objectives are:

- To define the lending decision task and its constraints in a way that supports model-to-policy translation;
- To present a detailed and technically grounded methodology for training and validating a GBDT risk model using leakage-aware temporal splits;
- To show how calibration and policy thresholding change the operational meaning of model outputs; and
- The discussion will focus on governance, fairness tensions, and the limitations that arise from using boosted models in regulated consumer credit.

The remainder is organized as follows. Section 4 reviews prior research and practice across credit scoring, tree ensembles, calibration, cost-sensitive decisioning, and explainability. Section 5 defines the problem precisely and describes a decision pipeline. Section 6 details the proposed methodology and operational controls. Section 7 presents hypothetical results and discusses implications. Sections 8 and 9 conclude and outline future research directions.

## 2. Literature Survey

### 2.1 Classical Credit Scoring, Scorecards, and Practical Constraints

Consumer credit scoring has long been treated as a supervised classification task under constraints such as class imbalance, selection bias, and shifting populations. Hand and Henley's review remain influential for its treatment of model choice and practical complications such as reject inference and population drift (Hand & Henley, 1997). Thomas (2000) and Thomas, Crook, and Edelman (2002) consolidated scorecard

practices that combine engineering heuristics with statistical modeling: binning, weight-of-evidence transformations, and policy overlays that embed business rules.

Benchmarking studies pushed the field beyond a "scorecards are enough" narrative. Baesens et al. (2003) compared multiple methods for credit scoring and argued that differences are real but sensitive to data and evaluation choices. Lessmann et al. (2015) revisited benchmarking with more modern learners and showed that machine learning methods often outperform classical approaches on discrimination, though the operational acceptability depends on stability and governance considerations. Brown and Mues (2012) examined imbalanced credit datasets and illustrated that algorithm behavior changes materially when default is rare, which is typical in retail portfolios.

The related literature explored richer data sources. Khandani, Kim, and Lo (2010) showed that transaction-derived features can improve consumer credit-risk modeling, but that richer feature spaces increase leakage risk and can embed socio-economic proxies that raise governance questions. The underlying point is not simply that "more data helps," but that feature provenance and stability matter as much as predictive power.

### 2.2 Tree Ensembles and Boosting for Structured Risk Data

Tree-based ensembles gained traction because they handle mixed feature types and complex interactions without heavy manual transformations. Random forests provided a strong baseline ensemble method and highlighted how aggregation improves generalization (Breiman, 2001). Boosting offered a different mechanism: sequentially refining predictions to correct prior errors. Friedman's gradient boosting formulation formalized boosting as iterative improvement under a chosen loss function, making it a conceptual foundation for modern GBDTs (Friedman, 2001). Earlier work on boosting, including AdaBoost, demonstrated the broader principle that ensembles of weak learners can be turned into strong predictors (Freund & Schapire, 1997). Friedman, Hastie, and Tibshirani (2000) connected boosting to additive logistic models, which is particularly relevant for default prediction tasks that naturally align with log-loss optimization.

By the mid-2010s, implementation advances turned GBDTs into practical production tools. XGBoost introduced regularized tree boosting and engineering techniques that made training fast and scalable (Chen & Guestrin, 2016). LightGBM emphasized efficient histogram-based training and sampling strategies for large tabular datasets (Ke et al., 2017). CatBoost addressed challenges with categorical features and prediction shift, which can appear in credit datasets with many categorical codes (Prokhorenkova et al., 2018). These systems made it feasible to train models frequently and explore larger feature sets, but they also expanded the governance problem: the easier it becomes to add features and complexity, the harder it becomes to justify and control model behavior.

### 2.3 Calibration and Probability Quality

In lending, model outputs must function as reliable probability estimates, not only rankings. Calibration has a long tradition in forecasting, including proper scoring rules such as the Brier score (Brier, 1950). In machine learning, calibration challenges became prominent as algorithms optimized for margins and ranking. Platt (1999) introduced sigmoid-based calibration for SVM outputs. Zadrozny and Elkan (2002) demonstrated that isotonic regression can transform model scores into well-calibrated probabilities, often outperforming parametric approaches when enough data is available. Niculescu-Mizil and Caruana (2005) showed that strong discrimination does not guarantee good calibration and that boosting can yield probability distortions unless corrected.

Credit risk also adds the complication of time. Calibration measured in a development window can drift as macro conditions change or as the lender's own approval policy alters the booked population. This is one reason validation regimes and PD monitoring have a prominent role in banking governance frameworks (BCBS, 2004).

### 2.4 Cost-Sensitive Decisioning, Imbalance, and Policy Thresholds

Unsecured lending decisions are cost-asymmetric: approving a future defaulter is typically far more expensive than rejecting a good borrower. Elkan (2001) argued that classification should be aligned with costs, not generic accuracy metrics. Related work on imbalanced learning, such as SMOTE, addressed minority-class scarcity (Chawla et al., 2002), but rebalancing can distort probability estimates if not handled with care. In lending, the safer pattern is often to keep training aligned to real base rates, then manage costs and class imbalance primarily through thresholding, policy segmentation, and calibrated probabilities.

In fintech and P2P lending contexts, researchers explicitly framed boosted models as decision-support tools. Serrano-Cinca et al. (2015) investigated determinants of default in P2P lending and highlighted selection effects created by platform policies. Xia et al. (2017) proposed cost-sensitive boosted approaches for loan evaluation, reflecting a broader shift from "pure prediction" to decisioning frameworks that integrate business constraints.

### 2.5 Explainability, Fairness, and Accountability

Explainability is central in credit because decisions often require customer-facing reasons and internal auditability. LIME provided local surrogate explanations for black-box models (Ribeiro et al., 2016). SHAP offered a theoretically grounded feature attribution approach, with efficient computation for tree ensembles (Lundberg & Lee, 2017). Strumbelj and Kononenko (2014) provided earlier foundations for feature contribution explanations.

Fairness research complicates decision-making because different fairness criteria cannot generally be satisfied simultaneously when base rates differ across groups. Hardt et al. (2016) discussed equality of opportunity. Kleinberg et al.

(2017) and Chouldechova (2017) showed incompatibilities between calibration and equalized error rates under differing base rates—an important warning for lenders who want both calibrated PDs and parity of certain error metrics. Barocas, Hardt, and Narayanan (2019) systematized fairness and accountability concerns, offering conceptual tools that are increasingly necessary for modern credit governance.

Synthesis. The pre-2020 literature supports the claim that GBDTs can outperform traditional scorecards on discrimination, but it also makes clear that lending demands more than AUC: probability calibration, selection bias management, stability controls, and explainability translation are the actual deployment bottlenecks.

## 3. Problem Definition

### 3.1 Task Statement

Given an unsecured loan applicant described by a feature set available at decision time (application data, bureau attributes, and optional behavioral or channel signals), estimate the applicant's default risk over a fixed horizon (for example, 12 months) and translate that risk estimate into a decision outcome such as approve, refer, or decline. When the product design allows, the decision may also include the offered interest rate, credit limit, or term.

### 3.2 Decision Pipeline

A decisioning system typically separates modeling from policy:

1) Feature construction using only information available at decision time
2) GBDT model producing a risk score or PD estimate
3) Calibration layer converting the model's raw output into a probability that behaves like the observed default frequency
4) Policy engine applying thresholds and business rules to produce the final decision (and, optionally, pricing/limits)
5) Monitoring for drift, stability, and fairness signals

### 3.3 Feature Categories and Risks

| Feature Category | Examples | Common Failure Modes |
|---|---|---|
| Application | income band, employment tenure, residence stability | misreporting, inconsistent documentation |
| Credit bureau | delinquency counts, utilization, tradeline age | bureau lag, thin-file volatility |
| Internal behavior (if existing customer) | repayment history, utilization trend | survivorship bias, policy feedback loops |
| Channel/device | acquisition channel, device stability proxies | proxy discrimination, instability across campaigns |
| Macro/regional | stress index, unemployment proxy | spurious correlation, overfitting to period effects |

### 3.4 What is "Risk Decisioning"? Means Here

"Risk decisioning" is treated as a mapping from calibrated default probability into operational actions that are cost-sensitive. The aim is not to maximize a single metric but to produce a stable and auditable decision boundary aligned to portfolio objectives.

## 4. Methodology and Approach

### 4.1 Data Definition and Leakage Controls

Outcome definition. A clear default label is essential. Retail credit often uses delinquency-based events (for example, reaching 90+ days past due) or charge-off status within a horizon such as 12 months. Whatever definition is chosen must be consistently applied across cohorts and carefully time-stamped.

Temporal splitting. Random train/test splitting is usually inappropriate for unsecured lending because it leaks future conditions and overstates performance. A time-based split (train on earlier cohorts, validate and test on later cohorts) better approximates deployment reality and exposes drift sensitivity.

Leakage checks. Leakage frequently hides in engineered features that accidentally incorporate post-decision updates. Practical controls include:

- Strict feature timestamping;
- Excluding any field that can be updated after the decision unless snapshotting is guaranteed;
- Building "as-of" feature views; and
- Auditing top features for suspiciously high predictive power that appears only with random splits.

### 4.2 Feature Engineering Choices that Matter in Credit

GBDTs reduce the need for manual binning, but credit modeling still benefits from domain-aware engineering:

- Stability-oriented transformations. For variables like utilization or delinquency counts, consider winsorization or capped scaling to prevent extreme outliers from dominating split logic.
- Missingness handling. Tree systems can treat missing values as a separate branch direction. In credit data, missingness can reflect thin-file status; it can also proxy socio-economic differences. Creating explicit missing indicators for major fields often improves auditability.
- Categorical treatment. High-cardinality categories (employer, channel, and device signatures) are risky: they can overfit, drift quickly, and encode proxy effects. If used, apply strong regularization and stability tests across time windows. CatBoost-style methods can help, but the governance conversation becomes more demanding (Prokhorenkova et al., 2018).

### 4.3 GBDT Model Training Protocol

A technically defensible credit-risk GBDT protocol typically includes:

- Loss and objective. Binary classification objective aligned to default probability.

- Regularization. Conservative learning rate, early stopping on validation loss, row/column subsampling, and minimum leaf constraints to prevent overly specific segments.
- Hyperparameter search. Random search or Bayesian optimization can be used, but the validation scheme must remain time-based. A practical compromise is a staged search: start broad, then refine around stable regions of the space (Bergstra et al., 2011).
- Model selection criteria. Do not select solely on AUC. Include calibration measures and stability checks across subperiods. DeLong-style comparison can help if the organization requires statistical confirmation for marginal AUC differences (DeLong et al., 1988), but decision impact should be assessed with portfolio metrics.

### 4.4 Calibration as a First-Class Component

Raw boosted outputs can be poorly calibrated even when discrimination is strong (Niculescu-Mizil & Caruana, 2005). In unsecured lending, this matters because pricing, provisioning, and limit assignment require probabilities that track observed default rates.

Two practical calibration approaches are widely used:
- Sigmoid calibration (Platt scaling) is typically more stable with modest calibration data (Platt, 1999).
- Isotonic regression, more flexible but more sensitive to sample size and noise (Zadrozny & Elkan, 2002).

Calibration should be learned on a holdout window that resembles deployment conditions. If the portfolio is drifting, periodic recalibration may be preferable to frequent full retraining, provided governance rules support that separation.

### 4.5 Policy Engine: From Calibrated PD to Decisions

A policy engine generally combines risk estimates with hard rules (eligibility, fraud checks, affordability constraints) and then applies PD thresholds to determine approve/ refer/ decline. Thresholds should reflect asymmetric costs and portfolio constraints rather than generic classifier thresholds.

In practice, lenders often use:
- Risk bands (deciles or PD buckets);
- Segment-specific rules (thin-file vs. thick-file); and
- Volume/capital constraints (approval caps, exposure limits).

A key methodological point is that thresholding should be evaluated on time-forward tests, since a small PD calibration drift can produce unexpectedly large volume changes near a threshold.

### 4.6 Explainability and Reason-Code Translation

Model explanations are necessary but not sufficient in lending. SHAP and LIME can identify features contributing to an individual score (Ribeiro et al., 2016; Lundberg & Lee, 2017). However, adverse-action reasons typically need stable, understandable categories rather than raw feature attributions.

A robust practice is to:
- Map features into reason groups (for example, “high revolving utilization,” “recent delinquency,” and “short credit history”);
- Enforce monotonic or constrained behavior for key groups when appropriate; and
- Monitor reason-code frequency shifts over time as a stability signal.

### 4.7 Monitoring and Model Risk Controls

A production-ready unsecured lending model typically requires monitoring beyond performance metrics:

- Population stability indices (PSI) for key variables and the model score;
- Calibration drift checks by PD band;
- Subpopulation performance to detect segment degradation;
- Fairness and disparate impact monitoring aligned to applicable laws and internal policy; and
- Champion–challenger governance with rollback capability.

## 5. Results and Discussion

### 5.1 Setup

Assume 1.2 million historical applications with observed 12-month outcomes for booked loans. The observed booked default rate is roughly 5%. The evaluation uses time-based splits and a stable feature snapshot.

Models:

- Logistic regression scorecard-style baseline
- GBDT risk model (regularized, early-stopped)
- Calibrated GBDT (isotonic calibration fitted on validation window)

### 5.2 Predictive Performance Summary

Hypothetical test-window outcomes:

| Model | ROC-AUC | KS | PR-AUC | Calibration quality (Brier trend) |
|---|---|---|---|---|
| Logistic regression | 0.74 | 0.38 | 0.2 | Strong baseline |
| GBDT (raw) | 0.79 | 0.46 | 0.26 | Slightly overconfident in tails |
| GBDT + calibration | 0.79 | 0.46 | 0.26 | Improved probability alignment |

Interpretation
- The GBDT improves separation, consistent with credit-scoring benchmarking literature that often finds gains from non-linear methods (Baesens et al., 2003; Lessmann et al., 2015).
- The raw GBDT can show probability distortion despite strong ranking, echoing prior findings on boosted model calibration (Niculescu-Mizil & Caruana, 2005).
- Post-hoc calibration improves probability usability without materially harming discrimination, aligning with calibration research (Zadrozny & Elkan, 2002).

### 5.3 Tail Separation and Operational Meaning

Credit policy often benefits when the riskiest band contains a higher share of eventual defaulters. Hypothetically:

- Top 10% riskiest applicants capture 28% of observed defaults under logistic regression.
- The calibrated GBDT captures 36% in the same top band.

This difference can translate into real operating choices: tighter verification for the highest band, more conservative exposure limits, and improved collections prioritization. The mechanism is not mysterious; GBDTs often learn interactions that scorecards miss unless analysts explicitly engineer them.

### 5.4 What Improves, and What Gets Harder

What improves:

- Interaction modeling becomes automatic (for example, utilization behaves differently for thin-file vs. established credit histories).
- Non-linear cut points emerge naturally without manual binning.
- Mixed feature types can be used with fewer transformations.

What gets harder:

- Explanation outputs become less stable, especially if many correlated features are present. Local attributions can shift even when the applicant looks “similar” to a policy analyst.
- Model drift becomes more operationally visible: a small shift in a key feature distribution can change the ensemble’s decision pathways and move volume around thresholds.
- Governance review becomes deeper, especially when alternative data or channel/device features are included.

### 5.5 Limitations and Failure Modes

Selection bias and reject inference. Outcomes are observed only for approved applicants. When policies change, the booked population changes, and the model learns a policy-conditioned reality (Hand & Henley, 1997; Serrano-Cinca et al., 2015). This is not a minor technicality; it can dominate observed improvements if not handled carefully.

Macro regime shifts. A model trained in benign credit conditions can degrade under economic stress. Banking governance frameworks emphasize ongoing validation and conservative controls for this reason (BCBS, 2004).
Fairness tensions. Fairness results show that calibration and certain parity constraints can conflict when base rates differ (Kleinberg et al., 2017; Chouldechova, 2017). In lending, a practical response is to define fairness evaluation criteria explicitly, then quantify trade-offs rather than assuming a single “fair” target exists.

Calibration drift. Calibration learned on one period can drift quickly; periodic recalibration can help but adds governance complexity (Zadrozny & Elkan, 2002).

### 5.6 Real-World Implications

If the hypothetical improvements hold, GBDT decisioning can support:

- More precise verification allocation (reduce cost while controlling risk);
- Clearer segmentation for pricing and limit policies, assuming regulatory and consumer fairness constraints are respected; and
- Earlier detection of portfolio risk concentration, improving risk management responsiveness.

A careful caveat belongs here: a model that improves AUC but destabilizes reason codes or fails stability checks may be unacceptable in consumer lending. The deployment decision is therefore a governance decision as much as a modeling decision.

## 6. Conclusion

GBDTs provide a strong modeling option for unsecured lending risk decisioning because they handle non-linearities and interactions in tabular credit data more naturally than classical scorecards. When built with leakage-aware temporal validation, conservative regularization, and a dedicated calibration layer, a GBDT system can deliver better tail separation and more effective risk segmentation while producing probability estimates usable for policy, pricing, and portfolio controls. The central challenge is not training the model, but controlling it: stability under drift, selection bias from approval policies, explanation-to-reason-code translation, and fairness monitoring are the practical fault lines. A decisioning-oriented pipeline that treats calibration and governance as core components offers a workable path to responsible deployment.

## 7. Future Scope

1) Modern reject inference compatible with GBDTs. Research is needed on robust methods that reduce policy-conditioning without relying on fragile assumptions.
2) Constrained boosting for governance. More empirical work on monotonic and interaction constraints in credit risk would clarify the performance–stability trade-off.
3) Drift-aware calibration. Linking recalibration strategies to macro indicators and cohort effects could improve probability reliability without frequent retraining.
4) Fairness-aware decisioning under real credit constraints. Practical frameworks that translate fairness theory into product-aligned monitoring and mitigations remain underdeveloped.
5) Reason-code stability as a measurable objective. Developing quantitative stability targets for explanations could bring interpretability into the same discipline as AUC and calibration metrics.

## References

[1] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. A. K., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*.
[3] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[4] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
[5] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
[6] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.
[7] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453.
[8] Basel Committee on Banking Supervision (BCBS). (2004). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework (Basel II)*. Bank for International Settlements.
[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
[10] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.
[11] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.
[12] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
[13] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845.
[14] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of ITCS*.
[15] Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of IJCAI*.
[16] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
[17] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.
[18] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.
[19] Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28*(2), 337–407.
[20] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic curve. *Radiology, 143*(1), 29–36.
[21] Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A, 160*(3), 523–541.
[22] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[23] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
[24] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
[25] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[26] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787.
[27] King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2), 137–163.
[28] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of ITCS*.
[29] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of IJCAI*.
[30] Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.
[31] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[32] Miller, R. G. (1981). *Simultaneous Statistical Inference* (2nd ed.). Springer.
[33] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of ICML*.
[34] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*.
[35] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[36] Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning, 42*(3), 203–231.
[37] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of KDD*.
[38] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432.
[39] Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLOS ONE, 10*(10), e0139427.
[40] Strumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems, 41*(3), 647–665.

[41] Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149–172.
[42] Thomas, L. C., Crook, J. N., & Edelman, D. B. (2002). *Credit Scoring and Its Applications*. SIAM.
[43] Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1).
[44] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann.
[45] Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using cost-sensitive learning for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications, 24*, 30–43.
[46] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of KDD*.