Model Performance, Fairness, and Policy Analysis for Home Credit Default Risk

A case study in AI for Finance: accuracy, explainability, and fairness

# Executive Summary

Credit risk modeling plays a critical role in ensuring financial stability by enabling lenders to distinguish between safe and risky borrowers. With the growing adoption of Artificial Intelligence (AI) in finance, predictive models have become more accurate but also more opaque, raising concerns around fairness, accountability, and regulatory oversight.

This project uses the Home Credit Default Risk dataset as a case study to explore not only how well machine learning models can predict default, but also whether they do so fairly across demographic groups. Unlike competition-driven approaches that optimize solely for predictive accuracy, this study explicitly integrates fairness evaluation and explainability into the modeling pipeline. The analysis aims to inform both financial institutions and regulators on the opportunities and risks of AI in credit scoring.

Methodology: The study merged multiple sources of customer and loan data; engineered financial and demographic features (e.g., credit utilization, repayment history, external credit scores); compared Logistic Regression (transparent baseline) with LightGBM (non-linear gradient boosting); evaluated models using AUC, accuracy, precision, and recall; applied SHAP for global and local explainability; and assessed fairness via selection rate, true positive rate (equal opportunity), and false positive rate across gender and age groups.

Key Findings: LightGBM marginally outperformed Logistic Regression (AUC 0.778 vs. 0.767). However, both models exhibited low precision (<20%), indicating many applicants flagged as high risk were actually creditworthy. External credit scores (EXT\_SOURCE\_1–3) and loan-related variables were the strongest predictors. Fairness analysis showed that male and younger applicants were flagged as high risk more frequently—leading to higher detection rates but also more false positives—while older applicants were flagged less often but their defaults were more likely to be missed.

Policy Implications: Accuracy alone is insufficient for assessing credit AI. Regulators should mandate fairness audits, standardize explainability (e.g., SHAP) for both oversight and consumer communication, strengthen data governance (especially around missing demographics), and require transparent justification of decision thresholds that shape approval and rejection rates. These recommendations align with global regulatory trends, including the EU AI Act’s treatment of credit scoring as a high-risk AI application.

Conclusion: Advanced algorithms can improve predictive power yet introduce disparities that affect financial inclusion. By combining predictive modeling with fairness and explainability analysis, this work demonstrates a framework that is technically robust and aligned with regulatory priorities for trustworthy AI in finance.

# Methodology & Results

The dataset provided by Home Credit comprises a wide range of tables describing applicants, their financial history, and behavioral attributes. To construct a reliable modeling base, we performed the following steps:

## Data Preparation

* Data Integration: Merged application data with auxiliary tables (bureau records, previous loan history, credit card balances).
* Data Cleaning: Addressed missing values through imputation while retaining demographic indicators for fairness analysis.
* Feature Engineering: Derived credit utilization ratios, repayment consistency, and loan-to-goods price ratios.
* Target Variable: Default encoded as a binary outcome (1 = default, 0 = successful repayment).

## Modeling Approach

* Logistic Regression: Interpretable baseline commonly used in financial risk modeling.
* LightGBM: Tree-based ensemble capturing non-linear interactions.
* Validation: Standard train/validation/test split with cross-validation for robustness.

## Evaluation Metrics

* AUC (Area Under ROC Curve) – ranking ability.
* Accuracy – overall correctness.
* Precision – correctness among predicted defaults (controls false alarms).
* Recall – coverage of actual defaults (controls missed risky borrowers).

## Explainability Methods

* Global SHAP to identify influential features across the dataset.
* Local SHAP explanations for representative applicants (high-risk, low-risk, borderline).

## Fairness Evaluation

* Dimensions: gender and age group.
* Metrics: selection rate, true positive rate (equal opportunity), false positive rate.
* Purpose: reveal demographic disparities relevant to regulatory oversight.

## Model Performance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AUC | Accuracy | Precision | Recall |
| Logistic Regression | 0.767 | 0.707 | 0.172 | 0.691 |
| LightGBM | 0.778 | 0.757 | 0.196 | 0.647 |

- LightGBM achieved marginally higher AUC and accuracy than Logistic Regression.  
- Both models exhibited low precision (<0.20), meaning many applicants flagged as high risk were actually creditworthy.  
- Recall values (~65–69%) indicate the models captured a reasonable share of true defaults.

## Global Feature Importance

The strongest predictors were external credit scores (EXT\_SOURCE\_1, EXT\_SOURCE\_2, EXT\_SOURCE\_3) and loan size variables (e.g., AMT\_GOODS\_PRICE). Some demographic indicators (e.g., CODE\_GENDER\_M) showed influence, warranting careful fairness scrutiny.

## Local Explanations

* High-risk applicant: Low external scores and high loan-to-goods ratio dominated the prediction.
* Low-risk applicant: Strong external scores and stable history reduced risk probability.
* Borderline applicant: Mixed signals (moderate external score but high credit utilization) led to uncertain classification.

## Fairness by Gender

Male applicants were flagged as high risk more often (selection rate ≈ 35.6% vs. 22.1% overall). They achieved higher true positive rates (≈ 73%), so defaults among men were detected more frequently; however, they also suffered higher false positive rates (≈ 31%), implying more creditworthy men were wrongly flagged. Records with missing gender displayed poor performance, indicating data quality issues.

## Fairness by Age Group

Younger applicants (30–40) experienced higher selection rates (~33%) with strong detection of defaults but also more false positives. Older applicants (60–70) were flagged much less (~12%), and their defaults were more likely to be missed (TPR ≈ 38%). This indicates a systematic bias: younger groups face more scrutiny, while some risky older borrowers may be under-detected.

## Interpretation of Results

The analysis highlights a trade-off between predictive performance and fairness. The models successfully leverage external credit scores, but demographic disparities persist. Younger and male borrowers are disproportionately classified as risky, potentially limiting equitable access to credit, whereas older borrowers may be under-scrutinized, elevating portfolio risk. Optimization must therefore incorporate fairness and explainability as first-class objectives.

# Policy Recommendations

The results of this study underscore that while advanced machine learning models can enhance credit risk prediction, they also create fairness concerns that directly affect financial inclusion. Governance frameworks must balance predictive accuracy with ethical and equitable practices.

1. Mandated Fairness Audits: Regulators should require regular evaluations across demographic groups with metrics such as selection rate, true positive rate, and false positive rate, reported alongside traditional performance measures.

2. Standardization of Explainability: Enforce explainable AI methods (e.g., SHAP) for regulatory transparency and consumer communication, reducing the opacity of black-box models.

3. Integration of Fairness-Aware Modeling: Encourage re-weighting, group-specific thresholds, or post-processing corrections to reduce disparate impacts while maintaining performance.

4. Strengthening Data Governance: Enforce minimum data quality standards, audit handling of missing values, and document data lineage and preprocessing to ensure end-to-end auditability.

5. Transparency in Thresholds and Risk Policies: Require institutions to justify decision thresholds, assess group impacts, and publish aggregate approval/rejection rates by group.

6. Alignment with Global Regulatory Trends: Harmonize with frameworks like the EU AI Act by defining minimum standards for risk, fairness, and transparency; require impact assessments before deployment; and promote sharing of best practices.

Conclusion: The deployment of AI in credit risk modeling offers improved predictive power but risks systemic bias. The framework demonstrated here integrates predictive modeling, fairness analysis, and explainability in a manner aligned with the practical governance needs of AI in finance.