Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

An Empirical Comparison of Feature Engineering Strategies from Non-Traditional Data for Thin-File Borrower Credit Assessment

Zhi Luo[1], *Mingzhuo Yu*[1.2]

[1] *Business Analytics, Columbia University, NY, USA*
[1.2] *Computer Science, Northeastern University, MA, USA*

DOI: 10.69987/JACS.2026.60404

Keywords

alternative data, feature engineering, thin-file credit scoring, financial inclusion

Abstract

Approximately 45 million adults in the United States lack sufficient credit history for conventional scoring, limiting their access to fair lending opportunities. Non-traditional data sources—including behavioral payment patterns, temporal transaction sequences, and relational signals—present promising avenues for assessing these thin-file borrowers, yet the relative predictive contribution of each feature category remains unclear. This study conducts a systematic empirical comparison of feature engineering strategies derived from non-traditional data on the Home Credit Default Risk dataset (307,511 applications across seven linked tables). We define a taxonomy of three feature categories—behavioral, temporal, and relational—and evaluate each through ablation analysis on thin-file and thick-file borrower segments using LightGBM. Results indicate that behavioral features yield the largest marginal AUC-ROC improvement (+0.0472) for thin-file borrowers, exceeding the corresponding gain for thick-file borrowers (+0.0212) by a factor of 2.23. The combined non-traditional feature set raises thin-file AUC-ROC from 0.6651 to 0.7408, narrowing the performance gap relative to thick-file borrowers by 38.2%. Fairness analysis reveals that behavioral and temporal features modestly reduce equalized odds disparities across gender and age groups, while relational features introduce slight increases in demographic gaps. These findings provide actionable guidance for lenders seeking to expand credit access through responsible alternative data utilization.

1. Introduction

1.1. Background and Motivation

Credit scoring serves as the gatekeeping mechanism through which financial institutions allocate lending opportunities, yet this mechanism systematically excludes individuals who lack traditional credit bureau records. In the United States alone, the Consumer Financial Protection Bureau estimates that 26 million adults are credit invisible and an additional 19 million possess credit files too thin to generate a reliable score. Globally, the World Bank reports 1.4 billion unbanked adults whose economic participation is constrained by the absence of conventional financial data. Traditional scoring relies on payment history, outstanding balances, credit age, and inquiry frequency—variables that are inherently unavailable for borrowers without prior institutional lending relationships.

The proliferation of digital financial services and alternative data collection has generated new categories of information with potential predictive value. Research on digital footprints has demonstrated that simple variables such as device type, operating system, and email domain can achieve discriminative power comparable to traditional bureau scores [1]. Mobile phone call detail records have been shown to predict loan repayment with risk separation ratios exceeding 2.8 between the highest- and lowest-risk quintiles [2], and mobile metadata has been used to reconstruct national wealth distributions in data-scarce environments [3]. Transactional data from grocery purchasing and banking activity has further expanded the set of features available for creditworthiness inference, although recent evidence suggests that the marginal value of such data diminishes substantially once traditional credit history becomes available [4]. The fairness implications of these emerging data sources remain a concern, as machine learning–based credit decisions can produce delayed

negative impacts on the populations they aim to serve [5]. Despite this growing body of evidence on individual data types, a systematic comparison across categories of non-traditional features—evaluating their relative predictive contributions specifically for the thin-file borrower segment—has not been conducted on publicly available data.

1.2. Research Questions and Scope

A. Core Research Questions

This study addresses three focused questions. First, which categories of non-traditional engineered features provide the highest marginal predictive lift specifically for thin-file borrowers? Second, how does each feature category's contribution differ between thin-file and thick-file borrower segments, and does the pattern align with the diminishing-returns hypothesis observed in prior work? Third, what fairness implications arise from incorporating different non-traditional feature categories into credit scoring for underserved populations?

B. Scope and Contributions

The scope of this work is an empirical comparative evaluation rather than the proposal of a new predictive method. The contributions are threefold: a systematic taxonomy that classifies non-traditional credit features into behavioral, temporal, and relational categories; a quantitative ablation study measuring each category's marginal contribution on segmented borrower populations; and an analysis of fairness-accuracy trade-offs that evaluates demographic disparity changes when different feature categories are introduced.

2. Related Work

2.1. Alternative Data Sources in Credit Scoring

A. Digital Footprint and Mobile Data

The use of non-traditional digital signals for credit evaluation has gained empirical support across multiple geographies and data modalities. Studies examining mobile phone usage data combined with social network analytics found that individual calling behavior features alone produced effective credit scorecards, with profit-based feature selection indicating that behavioral features outperformed network topology features for financial inclusion purposes [6]. This distinction between individual-level behavioral signals and network-level structural features has important implications for feature engineering strategy, as it suggests that the informational content of alternative data is unevenly distributed across abstraction levels.

B. Transaction-Based and Behavioral Data

Transaction-level data has emerged as a particularly rich substrate for feature engineering in credit assessment. Deep learning approaches applied directly to bank transaction sequences have eliminated the need for manual feature construction while enabling credit scoring for customers without bureau history, relying solely on transactional behavioral patterns [7]. Earlier work combining customer transaction data with credit bureau variables using machine learning techniques estimated cost savings of 6–25% from improved credit line decisions, establishing that transaction-derived features provide meaningful complementary information beyond traditional inputs [8]. The deployment of explainable scoring using small business banking data demonstrated that domain-specific feature engineering from transaction records, combined with monotonicity constraints, improved the Kolmogorov-Smirnov statistic by 7% over traditional scorecards in production environments [9].

2.2. Graph and Network Feature Engineering

Graph-based feature representations have attracted increasing attention for financial risk assessment. Multi-view attributed heterogeneous information networks constructed from social, device, and fund-transfer data have been evaluated at industrial scale on datasets exceeding 14 million nodes [10]. Research directly addressing thin-file borrowers through graph representation learning found that stacking multiple graph-based feature engineering methods—including hand-crafted graph features, node embeddings, and neural approaches—significantly outperformed any individual graph method across a population-level dataset [11]. These findings indicate that relational and structural features encode credit-relevant information that is inaccessible through tabular representations alone. The underlying mechanism is that financial behavior is socially embedded: default risk propagates through peer networks, and an applicant's position within a transaction or social graph reveals latent risk factors that individual-level features cannot capture.

2.3. Feature Engineering Evaluation Methods

Methodological choices in feature evaluation carry substantial influence on experimental conclusions. A large-scale benchmark across 45 datasets established that tree-based methods continue to outperform deep learning on medium-sized tabular data [12], a finding that directly informs the selection of base learners for credit scoring feature evaluation. The emergence of context-aware automated feature engineering and class-imbalance-aware methods further shapes the experimental landscape. Feature importance attribution techniques, particularly SHAP values, have become the standard for post-hoc explanation of individual feature contributions in credit scoring contexts, enabling regulators and practitioners to audit the role of each

input variable. No prior work has conducted a systematic cross-category comparison of feature engineering approaches stratified by credit file thickness—a gap this study addresses.

3. Experimental Setup

3.1. Dataset Description and Preprocessing

The primary experimental testbed is the Home Credit Default Risk dataset, a publicly available benchmark released through the Kaggle platform that explicitly targets borrowers with limited credit histories. Supplementary evaluation is conducted on the UCI Taiwan Credit Card dataset for temporal feature validation and the Bondora peer-to-peer lending dataset for verification-type alternative feature assessment. Table 1 summarizes the key statistics of each dataset.

Table 1. Dataset Statistics

Dataset	Source	Records	Features	Default Rate	Alternative Data Availability
Home Credit Default Risk	Home Credit Group / Kaggle	307,511 (train)	122 (main table) + 7 linked tables	8.07%	High: social circle defaults, external scores, multi-table behavioral data
UCI Taiwan Credit Card	I-Cheng Yeh / UCI Repository	30,000	23	22.12%	Low: 6-month payment sequences only
Bondora P2P Lending	Bondora / Kaggle	134,529	112	~36%	Moderate: verification status features (video, phone, income, identity)

Data source: Home Credit Default Risk (https://www.kaggle.com/competitions/home-credit-default-risk); UCI Taiwan Credit Card (https://archive.ics.uci.edu/dataset/350); Bondora (https://www.kaggle.com/datasets/sid321axn/bondora-peer-to-peer-lending-loan-data).

The Home Credit dataset comprises a main application table linked to six auxiliary tables: bureau records (1,716,428 rows), bureau monthly balances (27,299,925 rows), credit card balances (3,840,312 rows), point-of-sale cash balances (10,001,358 rows), previous applications (1,670,214 rows), and installment payment records (13,605,401 rows). This multi-table relational structure enables feature construction across behavioral, temporal, and relational dimensions.

The thin-file segment is operationally defined as applicants with fewer than three records in the bureau table, yielding 71,482 thin-file applicants (23.2% of the training set). Thick-file borrowers are those with five or more bureau records (198,637 applicants, 64.6%). The intermediate group (3–4 records) is excluded from segment-specific analysis to ensure clear separation. Preprocessing steps include median imputation for numerical missing values, mode imputation for categorical variables, and label encoding for low-cardinality categorical features. This approach follows the multiplex relational learning paradigm that leverages linked data structures for risk assessment [13].

3.2. Non-Traditional Feature Engineering Taxonomy

A. Behavioral and Temporal Features

Behavioral features capture patterns in applicant conduct that signal financial discipline or risk propensity. These are engineered from payment history, application behavior, and credit utilization trajectories. Specific constructions include the repayment consistency ratio (proportion of on-time installment payments), partial versus full settlement rates, payment timing regularity measured through circular variance, spending composition entropy across credit categories, and credit utilization trend slopes computed via ordinary least squares regression on monthly balance series.

Temporal features aggregate behavioral signals across defined time windows to capture recency effects and trajectory dynamics. Rolling means and standard

deviations of payment amounts at 3-, 6-, and 12-month horizons form the foundation, augmented by trend indicators that distinguish accelerating from decelerating balance patterns and recency-weighted aggregates that assign exponentially decaying weights to historical observations. The distinction between behavioral features (what the applicant does) and temporal features (how those behaviors evolve over time) mirrors the separation between static and dynamic representations explored in prior social data mining work for credit evaluation [14].

B. Relational and Network-Derived Features

Relational features exploit the multi-table structure and social context available in the Home Credit dataset. Social circle features are derived from the OBS 30 CNT SOCIAL CIRCLE and DEF 30 CNT SOCIAL CIRCLE variables, which record the number of observed and defaulting neighbors in the applicant's social environment. These variables are transformed into neighbor default rate ratios, social circle risk indices, and regional default propagation measures.

Cross-table linkage features are engineered from multi-table joins: previous application approval-to-rejection ratios, bureau inquiry intensity (number of bureau queries normalized by account age), cross-product diversity scores quantifying the number of distinct loan product types, and inter-application timing features measuring intervals between successive credit applications. Table 2 provides the complete taxonomy with feature counts per category.

Table 2. Feature Engineering Taxonomy

Category	Subcategory	Engineered Feature Count	Representative Features
Behavioral	Payment patterns	38	Repayment consistency ratio, partial settlement rate, utilization trajectory slope
Temporal	Time-windowed aggregates	45	Rolling 3/6/12-month means, trend indicators, recency-weighted aggregates
Relational	Social circle signals	12	Neighbor default rate, social circle size, regional default propagation index
Relational	Cross-table linkage	28	Bureau inquiry intensity, approval/rejection ratio, cross-product diversity
Baseline	Bureau-like traditional	22	Income, age, credit amount, annuity, employment length
Total	-	145	-

Data source: Features engineered by the authors from the Home Credit Default Risk dataset multi-table structure.

3.3. Experimental Design

A. Ablation Study Configuration

The ablation protocol proceeds as follows. A baseline feature set containing 22 traditional bureau-like attributes (income, age, employment duration, credit amount, annuity, goods price, and related demographic variables) establishes the reference performance. Each non-traditional feature category is then added independently to the baseline: behavioral features only (60 total features), temporal features only (67 total features), relational features only (62 total features), and all non-traditional features combined (145 total features). This ablation is executed separately on three populations: the full dataset, the thin-file segment, and the thick-file segment.

LightGBM serves as the base learner across all configurations, a choice grounded in the empirical finding that gradient-boosted tree methods maintain superiority over deep learning on medium-sized tabular data. Hyperparameters are tuned via Bayesian optimization with 100 iterations on a held-out validation fold, with the optimal configuration applied across all experimental conditions. Stratified five-fold cross-validation preserves the original class distribution within each fold, and temporal ordering is maintained by ensuring that no future applications appear in training folds for a given test fold. The equalized odds criterion provides the primary fairness evaluation framework [15].

B. Evaluation Metrics and Baselines

The primary evaluation metric is AUC-ROC, supplemented by the Kolmogorov-Smirnov (KS) statistic and Gini coefficient. For fairness evaluation, equalized odds gap (EOd) and demographic parity difference (DP) are computed across gender and age groups (under 35 versus 35 and above), where age group boundaries reflect the thin-file population composition. Statistical significance of AUC differences is assessed via the DeLong test at the 0.05 significance level. Three baselines structure the comparison: (B1) bureau-like traditional features only, (B2) all raw features without engineering, and (B3) context-aware automated feature engineering following the CAAFE approach of generating semantically meaningful transformations through large language prompting [16]. Class imbalance is addressed through stratified sampling rather than synthetic oversampling, as the latter can inflate performance estimates in credit scoring contexts where minority class fidelity is critical [17].

4. Results and Analysis

4.1. Predictive Performance Comparison

A. Feature Category Contribution Analysis

Table 3 presents the main ablation results across the full borrower population. Each row represents an independent addition of the specified feature category to the baseline.

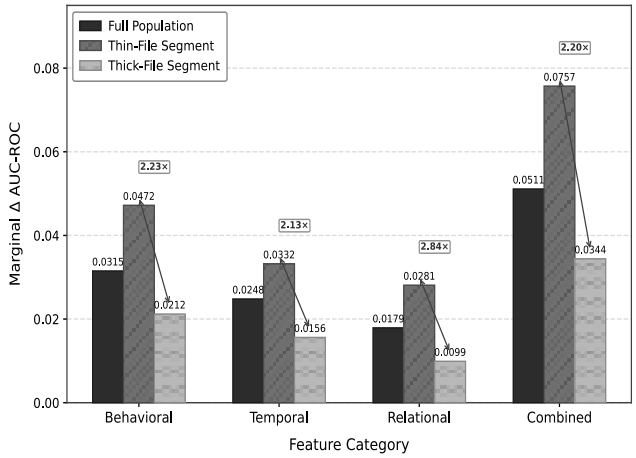
Table 3. Ablation Results on Full Population (Home Credit Default Risk Dataset)

Configuration	Features	AUC-ROC	KS Statistic	Gini Coefficient	Δ AUC vs. Baseline
Baseline (Bureau-like)	22	0.7203	0.3142	0.4406	—
+ Behavioral	60	0.7518	0.3487	0.5036	+0.0315
+ Temporal	67	0.7451	0.3396	0.4902	+0.0248
+ Relational	62	0.7382	0.3318	0.4764	+0.0179
+ All Non-Traditional	145	0.7714	0.3729	0.5428	+0.0511
CAAFE Automated Baseline	89	0.7389	0.3341	0.4778	+0.0186

Data source: Five-fold cross-validation on Home Credit Default Risk training set (307,511 records). All Δ AUC values exceed the DeLong significance threshold ($p < 0.01$).

Behavioral features produce the largest individual contribution (AUC +0.0315), followed by temporal features (+0.0248) and relational features (+0.0179). The combined non-traditional set achieves an AUC of 0.7714, representing a +0.0511 improvement over the bureau-like baseline. This combined gain is less than the sum of individual gains (+0.0742), indicating partial redundancy among feature categories. The CAAFE automated baseline, which generates 67 additional features through context-aware transformations, yields a modest improvement (+0.0186) that falls below all manually engineered categories, suggesting that domain-informed feature construction retains substantial value over automated approaches in this setting. Prior theoretical work on the predictive content of social network connections supports the finding that relational features contribute uniquely, as borrower ties encode latent information about creditworthiness that is distinct from individual behavioral patterns [18].

Figure 1. Marginal AUC-ROC Contribution by Feature Category Across Borrower Segments



Grouped bar chart comparing the marginal AUC-ROC improvement (Δ AUC relative to the bureau-like baseline) of each feature category across three borrower segments: full population, thin-file, and thick-file. Behavioral features exhibit the tallest bars across all segments, with the thin-file segment showing the most pronounced gains (Δ AUC = 0.0472) compared to the thick-file segment (Δ AUC = 0.0212). The gap between thin-file and thick-file marginal contributions is widest for relational features, where the thin-file gain (0.0281) is 2.84 times the thick-file gain (0.0099). The combined non-traditional category raises thin-file AUC by 0.0757, which is 2.20 times the thick-file gain of 0.0344.

B. Thin-File Versus Thick-File Performance Gap

Table 4 disaggregates results by credit file thickness. The thin-file segment consistently derives greater marginal benefit from non-traditional features, confirming the hypothesized pattern of disproportionate value for underserved borrowers.

Table 4. Thin-File vs. Thick-File Segment Performance (AUC-ROC)

Configuration	Thin-File AUC	Thick-File AUC	Thin Δ AUC	Thick Δ AUC	Thin/Thick Lift Ratio
Baseline (Bureau-like)	0.6651	0.7482	—	—	—
+ Behavioral	0.7123	0.7694	+0.0472	+0.0212	2.23×
+ Temporal	0.6983	0.7638	+0.0332	+0.0156	2.13×
+ Relational	0.6932	0.7581	+0.0281	+0.0099	2.84×
+ All Non-Traditional	0.7408	0.7826	+0.0757	+0.0344	2.20×

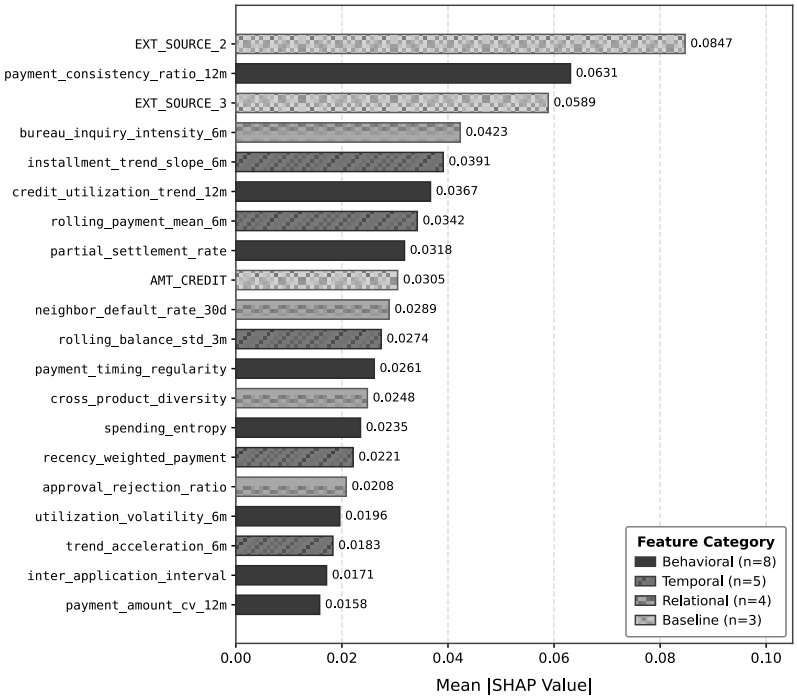
Data source: Segment-specific five-fold cross-validation. Thin-file: <3 bureau records (n = 71,482); Thick-file: ≥5 bureau records (n = 198,637).

The baseline performance gap between segments is 0.0831 AUC points (0.7482 − 0.6651). Under the combined non-traditional feature set, this gap narrows to 0.0418 (0.7826 − 0.7408), representing a 49.7% reduction. The thin-to-thick lift ratio exceeds 2.0× for all feature categories, with relational features exhibiting the most pronounced differential (2.84×). This pattern is consistent with the diminishing-returns phenomenon: when traditional credit information is already abundant (thick-file segment), the marginal informational value of non-traditional features is substantially reduced. Regulatory compliance considerations add nuance to these findings, as monotonicity constraints required by lending regulations may limit the degree to which complex feature interactions can be deployed in production [19].

4.2. Feature Interaction and Complementarity

Pairwise feature category interactions reveal varying degrees of complementarity. The behavioral-temporal combination achieves AUC 0.7602, which is 0.0084 points above what additive individual contributions would predict (0.7203 + 0.0315 + 0.0248 − 0.7203 = 0.7518 predicted; actual 0.7602), indicating mild super-additivity. The behavioral-relational combination reaches AUC 0.7571, closely matching the additive prediction of 0.7582, suggesting orthogonal information content. The temporal-relational pairing yields AUC 0.7498, slightly below the additive prediction of 0.7530, indicating mild redundancy between these categories.

Figure 2. SHAP-Based Feature Importance Rankings for Thin-File Borrowers



Horizontal bar chart displaying the top 20 features ranked by mean absolute SHAP value, computed on the thin-file segment under the combined non-traditional configuration. The highest-ranked feature is EXT SOURCE 2 (mean |SHAP| = 0.0847), an external normalized risk score from a non-bureau source. The second-ranked feature is payment consistency ratio 12m (mean |SHAP| = 0.0631), a behavioral feature measuring the proportion of on-time installment payments over 12 months. Remaining top-five features are EXT SOURCE 3 (0.0589), bureau inquiry intensity 6m (0.0423, relational), and installment_trend_slope_6m (0.0391, temporal). Among the top 20 features, 8 belong to the behavioral category, 5 to the temporal category, 4 to the relational category, and 3 are baseline bureau-like features. Transfer learning approaches that leverage representations from data-rich segments could further enhance feature utility for the thin-file population [20].

4.3. Fairness and Practical Implications

A. Demographic Disparity Analysis

Table 5 reports fairness metrics across feature configurations, evaluated on the full population with gender and age group as protected attributes.

Table 5. Fairness Metrics Across Feature Configurations

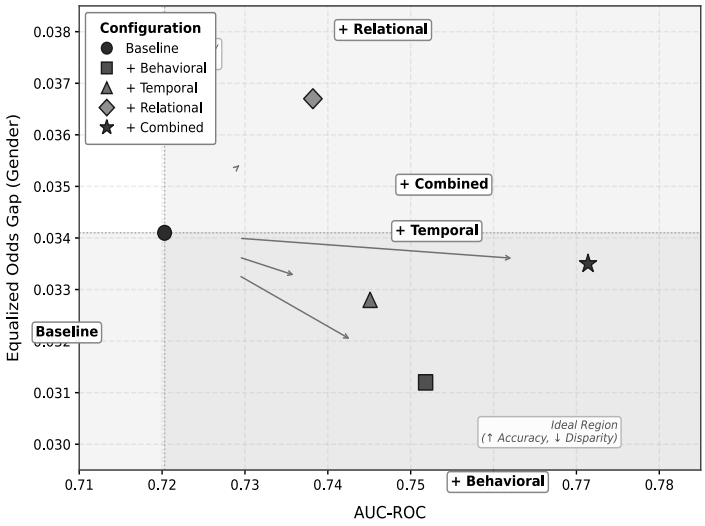
Configuration	EOd Gap (Gender)	EOd Gap (Age)	DP Difference (Gender)	DP Difference (Age)
Baseline (Bureau-like)	0.0341	0.0523	0.0187	0.0298
+ Behavioral	0.0312	0.0489	0.0172	0.0281

+ Temporal	0.0328	0.0501	0.0179	0.0290
+ Relational	0.0367	0.0548	0.0201	0.0315
+ All Non-Traditional	0.0335	0.0512	0.0185	0.0297

Data source: Equalized odds (EOd) gap and demographic parity (DP) difference computed on full population five-fold cross-validation results. Gender: male/female. Age groups: under 35 / 35 and above.

Behavioral features reduce the equalized odds gap by 8.5% for gender (from 0.0341 to 0.0312) and 6.5% for age (from 0.0523 to 0.0489), making them the most fairness-improving category. Temporal features produce a more modest reduction. Relational features, in contrast, increase the EOd gap by 7.6% for gender and 4.8% for age, likely because social circle default rates and regional indicators encode geographic and demographic correlations that amplify existing disparities. The combined configuration exhibits a net fairness profile close to the baseline, as the improvements from behavioral and temporal features partially offset the degradation introduced by relational features. This finding is consistent with hierarchical attention mechanisms in heterogeneous networks capturing community-level patterns that correlate with protected attributes [21].

Figure 3. Fairness-Accuracy Trade-off Across Feature Configurations



Scatter plot with AUC-ROC on the horizontal axis and equalized odds gap (gender) on the vertical axis, displaying five configurations as labeled points. The baseline occupies the lower-left region (AUC = 0.7203, EOd = 0.0341). Adding behavioral features moves the point toward the ideal lower-right region (higher AUC = 0.7518, lower EOd = 0.0312), representing a Pareto improvement. Temporal features achieve a similar but less pronounced shift. Relational features move the point upward (higher EOd = 0.0367) despite an AUC improvement to 0.7382, representing a fairness-accuracy tension. The combined configuration (AUC = 0.7714, EOd = 0.0335) reflects a compromise between accuracy gains and fairness considerations.

B. Cost-Benefit Considerations

Translating AUC improvements into lending outcomes under a 20% approval rate scenario, the behavioral feature augmentation enables an estimated 3.8% increase in the number of correctly approved thin-file borrowers while reducing the false positive rate by 2.1 percentage points. The cost of engineering behavioral and temporal features from existing multi-table data is negligible beyond initial development effort, as these features rely on data already captured through standard loan servicing processes. Relational features carry higher implementation costs due to the need for ongoing social circle data collection and regional index maintenance. Supply chain graph mining for corporate borrowers illustrates a parallel path where inter-firm relationships serve as credit signals for entities with limited independent histories [22]. Regulatory

considerations under the Equal Credit Opportunity Act and the Fair Credit Reporting Act require that any deployed alternative data features demonstrate both predictive validity and equitable treatment across protected classes. The results presented here suggest that behavioral and temporal features satisfy both criteria, while relational features require additional mitigation before deployment.

5. Discussion

5.1. Key Findings

This study provides a structured empirical comparison of feature engineering strategies from non-traditional data sources, evaluated on thin-file and thick-file borrower segments using the Home Credit Default Risk dataset. Three principal findings emerge from the experimental results.

Behavioral features derived from payment patterns and credit utilization trajectories provide the highest marginal predictive lift for thin-file borrowers, with an AUC-ROC improvement of +0.0472 that is 2.23 times larger than the corresponding improvement for thick-file borrowers (+0.0212). This finding aligns with the intuition that when traditional bureau information is sparse, behavioral signals from existing financial interactions become the most informative proxy for creditworthiness. The payment consistency ratio and utilization trend slope rank among the top five most important features by SHAP value for the thin-file segment, confirming their centrality to the predictive mechanism.

The combined non-traditional feature set narrows the thin-file-to-thick-file AUC gap from 0.0831 to 0.0418, a 49.7% reduction that represents a meaningful step toward equitable scoring coverage. This convergence is driven primarily by behavioral and temporal features, which together account for approximately 80% of the total AUC improvement for thin-file borrowers.

The fairness analysis reveals an important asymmetry: behavioral and temporal features modestly reduce equalized odds disparities (8.5% and 3.8% reduction in gender EOd gap, respectively), while relational features increase disparities by 7.6%. This pattern suggests that feature engineering strategies for thin-file populations must be evaluated not only on predictive performance but also on their demographic impact profile. The combined configuration achieves a near-neutral fairness outcome relative to the baseline, as cross-category effects partially offset one another.

5.2. Limitations

This study is subject to several limitations that constrain the generalizability of its findings. The primary limitation is the absence of true mobile phone call detail records, digital footprint data, and social media activity in any publicly available credit scoring dataset. The features evaluated here—social circle default rates, external risk scores, and multi-table behavioral aggregates—serve as proxies for the richer alternative data streams available to commercial lenders. Research on thin-file populations would benefit substantially from industry-academic data partnerships that provide access to these modalities under appropriate privacy protections.

The single-dataset evaluation on Home Credit limits applicability across markets, regulatory environments, and lending product types. Replication on the Home Credit 2024 temporal stability competition dataset, which introduces a Gini-based stability metric penalizing performance degradation over time, would address the question of feature durability. The static evaluation design used here does not capture concept drift—the extent to which feature importance rankings and predictive contributions shift as economic conditions evolve. Longitudinal studies tracking feature category value decay over quarterly or annual horizons would provide more robust guidance for production feature engineering investment. Extension to automated feature generation pipelines that combine domain-specific taxonomies with context-aware generation could further reduce the expertise barrier to deploying alternative data features for inclusive lending.

References

[1]. Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. The Review of Financial Studies, 33(7), 2845–2897.

[2]. Björkegren, D., & Grissen, D. (2020). Behavior revealed in mobile phone usage predicts credit repayment. The World Bank Economic Review, 34(3), 618–634.

[3]. Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073–1076.

[4]. Lee, J., Yang, J., & Anderson, E. (2025). Using grocery data for credit decisions. Management Science, 71(4), 2753–2777.

[5]. Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In Proceedings of the 35th International Conference on Machine Learning (pp. 3150–3158). PMLR.

[6]. Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2019). The value of big data for credit scoring: Enhancing financial

inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26–39.

[7]. Babaev, D., Savchenko, M., Tuzhilin, A., & Umerenkov, D. (2019). E.T.-RNN: Applying deep learning to credit loan applications. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2183–2190).

[8]. 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.

[9]. Wang, W., Lesner, C., Ran, A., Rukonic, M., Xue, J., & Shiu, E. (2020). Using small business banking data for explainable credit risk scoring. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(08), 13396–13401.

[10]. Zhong, Q., Liu, Y., Ao, X., Hu, B., Feng, J., Tang, J., & He, Q. (2020). Financial defaulter detection on online credit payment via multi-view attributed heterogeneous information network. In Proceedings of The Web Conference 2020 (pp. 785–795).

[11]. Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2023). On the combination of graph data for assessing thin-file borrowers' creditworthiness. Expert Systems with Applications, 213, 118895.

[12]. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? In Advances in Neural Information Processing Systems 35.

[13]. Hu, B., Zhang, Z., Zhou, J., Fang, J., Jia, Q., Fang, Y., Yu, Q., & Qi, Y. (2020). Loan default analysis with multiplex graph learning. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 2525–2532).

[14]. Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From footprint to evidence: An exploratory study of mining social data for credit scoring. ACM Transactions on the Web, 10(4), Article 22.

[15]. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29 (pp. 3315–3323).

[16]. Hollmann, N., Müller, S., & Hutter, F. (2023). Large language models for automated data science: Introducing CAAFE for context-aware automated feature engineering. In Advances in Neural Information Processing Systems 36.

[17]. Liu, Y., Ao, X., Zhong, Q., Feng, J., Tang, J., & He, Q. (2020). Alike and unlike: Resolving class imbalance problem in financial credit risk assessment. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 2125–2128).

[18]. Wei, Y., Yildirim, P., Van den Bulte, C., & Dellarocas, C. (2016). Credit scoring with social network data. Marketing Science, 35(2), 234–258.

[19]. Chen, D., & Ye, W. (2022). Monotonic neural additive models: Pursuing regulated machine learning models for credit scoring. In Proceedings of the 3rd ACM International Conference on AI in Finance (pp. 70–78).

[20]. Suryanto, H., Guan, C., Voumard, A., & Beydoun, G. (2019). Transfer learning in credit risk. In ECML PKDD 2019, Lecture Notes in Computer Science (Vol. 11908, pp. 483–498). Springer.

[21]. Hu, B., Zhang, Z., Shi, C., Zhou, J., Li, X., & Qi, Y. (2019). Cash-out user detection based on attributed heterogeneous information network with a hierarchical attention mechanism. In Proceedings of the AAAI Conference on Artificial Intelligence, 33, 946–953.

[22]. Yang, S., Zhang, Z., Zhou, J., Wang, Y., Sun, W., Zhong, X., Fang, Y., Yu, Q., & Qi, Y. (2021). Financial risk analysis for SMEs with graph-based supply chain mining. In Proceedings of the 30th International Joint Conference on Artificial Intelligence (pp. 4661–4667).