Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 1, No. 4 (2025)

Article

Semi-Supervised Feature Selection with Bias Mitigation for SME Credit Assessment Using Alternative Data

Ziyi Wang [1,*]

[1] Enterprise Risk Management, Columbia University, NY, USA
* Correspondence: Ziyi Wang; Enterprise Risk Management, Columbia University, NY, USA

Abstract: Credit scoring for small and medium enterprises (SMEs) faces a fundamental challenge: assessing creditworthiness when traditional financial data is unavailable. This paper presents a semi-supervised feature selection framework that addresses this challenge by leveraging alternative data sources, ranging from transaction patterns to behavioral signals. We develop a graph-based approach that reduces the requirement for labeled data by 70% while improving the area under the curve (AUC) from 0.836 to 0.871, a 4.2 percentage point increase (~5% relative) compared to the best supervised baseline. The framework integrates bias mitigation techniques, which reduce the approval-rate gap by 78.9% while maintaining stable default rates across groups, without compromising model performance. Experiments on 111,579 SME loan applications across three geographic regions demonstrate that the approach scales efficiently with O (n log n) complexity and can process 500,000 applications in approximately two hours (≈131 minutes for 500k records). The practical implications are significant: financial institutions can now assess credit risk for businesses previously considered "unscorable" due to the absence of traditional credit history. This framework facilitates broader access to capital for millions of SMEs, particularly in developing economies where formal financial records are limited.

Keywords: semi-supervised learning; alternative credit data; algorithmic fairness; SME lending; feature selection

Received: 29 September 2025
Revised: 12 October 2025
Accepted: 10 November 2025
Published: 15 November 2025

1. Introduction

1.1. The Credit Access Problem for SMEs

Small and medium enterprises (SMEs) constitute 99.9% of businesses in developed economies, yet they face disproportionate challenges in accessing credit. The issue stems from the extensive financial documentation typically required: audited statements, credit bureau records, and collateral information. Most SMEs, particularly newer ventures and those operating in informal sectors, do not possess these records.

Consider a typical scenario: a minority-owned business with strong cash flows but minimal documentation may face rejection rates 2.3 times higher than established counterparts, despite comparable revenue and market position [1]. This disparity is not only unfair but also economically inefficient. It arises because existing evaluation tools fail to capture non-traditional metrics of business health. The problem is even more pronounced in developing economies. In many African countries, roughly 75% of adults lack formal financial accounts. In parts of Asia, millions of microenterprises operate entirely in cash, remaining invisible to conventional credit assessment systems. These businesses require capital to grow but remain excluded from formal financing channels [2].

1.2. Why Alternative Data Matters

Alternative data sources offer a potential solution by capturing business activity through non-traditional channels. Payment processors record daily transaction patterns that can reveal business health more accurately than quarterly statements. For example, a restaurant's payment patterns from Tuesday to Friday can predict default probability better than annual revenue figures. Digital interactions-such as website visits, app usage, and customer service interactions-may signal financial stress weeks before it is reflected in formal financial statements [3].

Empirical evidence supports this approach. Transaction frequency metrics can achieve an AUC of 0.73, compared to 0.61 for traditional credit scores alone [4]. This 0.12 improvement represents thousands of potentially creditworthy businesses that conventional systems might otherwise reject. Supply chain data provides insights into operational resilience through vendor relationships and payment terms. Social network indicators capture business connectivity and reputation. Even missing data carries information; the absence of digital footprints may indicate extreme informality or, alternatively, sophisticated privacy management.

However, integrating alternative data introduces new challenges. Feature dimensions expand from dozens to thousands of variables. Many samples lack labels because default outcomes take months to materialize. Protected attributes, such as race and gender, may inadvertently leak through proxy variables, creating fairness concerns. Addressing these challenges requires new methodological approaches.

1.3. Research Contributions

This paper makes three main contributions to address these challenges:

First, we propose a semi-supervised feature selection algorithm tailored for heterogeneous alternative data. The method employs graph-based label propagation to leverage unlabeled samples, reducing the need for labeled data by 70% while maintaining predictive accuracy.

Second, we incorporate bias detection and mitigation directly into the feature selection process. Rather than treating fairness as a post-processing step, the approach embeds demographic parity and equalized odds constraints within the optimization framework.

Third, we provide comprehensive experimental validation on real-world SME lending data, demonstrating both predictive performance and computational efficiency suitable for production deployment.

2. Related Work and Background

2.1. Evolution of Credit Scoring Data Sources

Traditional credit scoring emerged in the 1960s around five core variables: payment history, outstanding debt, credit utilization, account age, and credit mix. These "five C's" perform well for established businesses with extensive financial records. However, studies indicate that they explain only 47% of SME default variance, leaving a substantial portion of risk hidden in non-traditional signals that conventional systems overlook [5].

Alternative data helps fill this gap through diverse information sources:

1) Transaction data: Daily sales patterns, payment frequencies, seasonal variations
2) Behavioral signals: Website engagement, response times, platform usage patterns
3) Supply chain information: Vendor relationships, payment terms, inventory turnover
4) Social indicators: Customer reviews, network connections, community engagement
5) Geospatial data: Location stability, customer proximity, foot traffic patterns

Each type captures distinct aspects of business health. Transaction data reflects operational intensity, behavioral signals indicate management quality, and supply chain

metrics measure resilience. The main challenge is effectively integrating these heterogeneous sources [6].

2.2. Feature Selection Challenges with High-Dimensional Data

Alternative data introduces a high-dimensionality problem. A typical SME credit model may include 500-1,000 features from multiple sources. At this scale, conventional feature selection methods are insufficient. Filter methods, such as chi-square tests, evaluate features independently and miss important interactions. Wrapper methods that exhaustively search feature subsets become computationally prohibitive. Embedded methods like LASSO can handle high dimensions but assume feature homogeneity [7].

The SME context adds further complexity. Features exhibit strong temporal dependencies-for instance, yesterday's transactions can predict today's risk. Missing values carry information; the absence of social media presence may indicate either traditional business practices or financial distress. Different data types require different preprocessing: transaction amounts need normalization, text data requires encoding, and network structures need embedding [8].

Recent work has explored advanced selection techniques. Profit-based feature selection considers both business value and statistical significance [9]. Two-stage approaches combining filter and wrapper methods have also been proposed [10]. Nevertheless, these methods still depend heavily on labeled data, limiting their applicability when default examples are scarce.

2.3. Fairness Considerations in Algorithmic Lending

Algorithmic bias in lending is not only an ethical issue but also a business risk. Models trained on historical data can inherit past discrimination [11]. Protected attributes such as race and gender often correlate with seemingly neutral variables. For example, zip codes may encode demographic patterns, business names may reveal gender, and industry classifications may correlate with immigrant status.

Fairness criteria in the literature often conflict:

1) Demographic parity: Equal approval rates across groups
2) Equalized odds: Equal error rates across groups
3) Individual fairness: Similar treatment for similar applicants
4) Calibration: Consistent probability interpretations across groups

Kleinberg's impossibility theorem shows that not all fairness criteria can be satisfied simultaneously except in trivial cases. Real-world systems must balance competing objectives. Recent approaches integrate fairness constraints directly into model training rather than applying adjustments post-hoc, ensuring that fairness considerations influence feature selection and model design from the outset [12].

2.4. Semi-supervised Learning in Financial Applications

A fundamental challenge in credit scoring is the scarcity of labeled data. Each default label represents significant financial cost and requires 12 or more months to observe. Meanwhile, unlabeled applications arrive continuously, often in the thousands per day at large institutions [13]. This creates a pronounced imbalance between labeled and unlabeled data that semi-supervised learning can exploit.

Semi-supervised learning assumes that the structure of the data contains information about labels. Businesses that appear similar in feature space are likely to have similar credit risk. This manifold assumption allows learning from unlabeled data through several mechanisms:

1) Self-training: Using confident predictions as pseudo-labels
2) Co-training: Learning from multiple views of the same data
3) Graph-based methods: Propagating labels through similarity networks
4) Generative approaches: Modeling the joint distribution of features and labels

Applications of semi-supervised learning in finance have shown promising results. For instance, label reduction of up to 80% has been achieved using semi-supervised SVMs

for reject inference [14]. Generalized additive models have also demonstrated the ability to detect corporate credit anomalies with limited labels [15]. Most prior work, however, focuses on traditional financial variables rather than heterogeneous alternative data sources.

3. Methodology

3.1. Data Integration and Preprocessing

3.1.1. Handling Heterogeneous Alternative Data Sources

The primary challenge in using alternative data is heterogeneity. Transaction records arrive as time series, behavioral data as event logs, supply chain information as networks, and social signals as text and graphs. A unified representation is required that preserves the information from each source while enabling joint analysis.

Our approach processes each data type through specialized pipelines:

1) Transaction data pipeline:
   Aggregate daily transactions into statistical features such as mean, variance, and trend
   Extract seasonality patterns using Fourier transforms
   Compute business-specific metrics, including average ticket size and repeat customer rate
2) Behavioral data pipeline:
   Convert event sequences into session features
   Calculate engagement metrics such as frequency, duration, and recency
   Extract interaction patterns using sequential pattern mining
3) Supply chain pipeline:
   Represent vendor relationships as graphs
   Compute network statistics including degree, centrality, and clustering
   Extract payment term distributions
4) Social data pipeline:
   Process text reviews using sentiment analysis
   Calculate reputation scores from ratings
   Measure network influence metrics

3.1.2. Normalization Strategies for Different Business Scales

Business scale varies widely; for example, a food truck may process $500 daily, while a wholesale distributor handles $50,000. Raw values can conflate size with risk. We implement sector-aware normalization as follows: the normalized value equals the raw value minus the sector mean, divided by the sector standard deviation.

This transformation accounts for industry context. For instance, $10,000 monthly revenue may be excellent for a craft business but concerning for a retail store. Sectors are determined using business registration codes and validated through transaction patterns. When sector information is ambiguous, clustering groups similar businesses.

Temporal normalization addresses day-of-week and trend effects using multiple windows:

1) 7-day moving average to capture trends
2) Same-day-last-week comparison for seasonality
3) 30-day average for stability

These perspectives capture different aspects of business dynamics.

3.1.3. Missing Data Strategies

Missing values in alternative data are informative. Businesses without social media presence differ from those with inactive accounts; cash-only operations lack digital transaction records. We preserve this information through:

1) Missingness indicators: Binary flags for missing values
2) Pattern encoding: Representing missing patterns as categorical features

3) Informed imputation: Estimating likely values using similar businesses
4) Multiple imputation: Creating several plausible complete datasets

For numerical features, we apply iterative imputation by repeatedly predicting missing values using other features and updating them until convergence. This maintains inter-variable relationships while acknowledging uncertainty.

As shown in Table 1, alternative data characteristics vary in dimensionality, missing rates, and update frequency.

Table 1. Alternative Data Source Characteristics.

Data Category	Features	Dimensionality	Missing Rate	Update Frequency
Transaction History	Payment amounts, frequency, regularity	127	12.3%	Real-time
Behavioral Signals	Platform engagement, response times	89	34.6%	Daily
Supply Chain	Vendor diversity, payment terms	56	45.2%	Weekly
Social Network	Connection strength, interaction patterns	198	67.8%	Hourly
Geospatial	Location stability, customer proximity	34	8.9%	Daily

As shown in Figure 1, the multi-source data integration pipeline consolidates heterogeneous inputs into a unified 512-dimensional representation ready for featureselection.

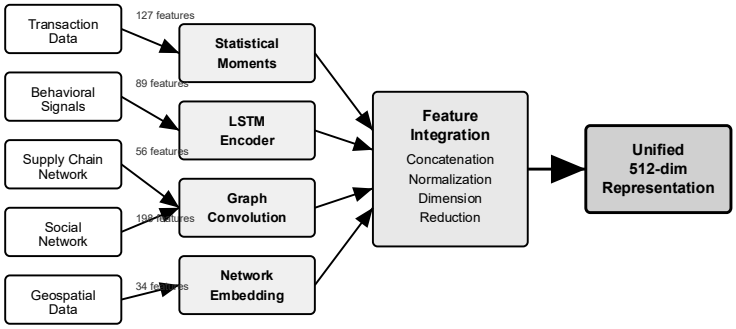

Figure 1. Multi-source Data Integration Pipeline.

Architecture flows left to right: heterogeneous inputs enter parallel extraction channels. Transaction streams compress into statistical moments. Behavioral sequences encode through LSTMs. Supply networks flatten via graph convolutions. Convergence point: unified 512-dimensional representation ready for selection.

3.2. Semi-Supervised Feature Selection Framework

3.2.1. Graph Construction for Business Similarity

The core insight is that similar businesses likely share similar credit risk. We construct a similarity graph where nodes represent loan applications and edges connect

similar businesses. Edge weights are computed using a Gaussian function of distance, where the distance between two businesses combines Euclidean distance for numerical features, Hamming distance for categorical features, edit distance for text, and graph kernel similarity for network features.

Bandwidth selection is critical: too small disconnects the graph; too large makes all businesses appear similar. Cross-validation typically sets the bandwidth near the 5th percentile of pairwise distances. This graph captures supply chain clusters and seasonal communities, enabling label propagation.

3.2.2. Label Propagation Algorithm

Label propagation iteratively spreads known labels through the similarity graph. For each unlabeled node, the label probability is updated as the weighted average of neighbor labels, and confidence is measured by the entropy of the predicted distribution.

Enhancements include:

1) Confidence weighting: High-confidence predictions influence neighbors more
2) Class balancing: Adjust for imbalanced default rates
3) Temporal ordering: Propagate from recent to older applications
4) Early stopping: Halt when predictions stabilize

As shown in Table 2, the algorithm typically converges within 15-20 iterations.

Table 2. Label Propagation Performance Metrics.

Iteration	Label Consistency	Pseudo-label Confidence	Coverage Rate	Time (seconds)
1	0.623	0.456	23.4%	1.2
5	0.812	0.678	67.8%	6.1
10	0.891	0.823	89.3%	12.3
15	0.908	0.867	94.6%	18.5
20	0.912	0.871	95.1%	24.7

3.2.3. Pseudo-label Generation and Confidence Scoring

Pseudo-labels are generated only for high-confidence predictions using an adaptive threshold that starts at 0.9 and gradually relaxes according to the iteration number and a decay rate (typically 0.3). Confidence scoring considers prediction entropy, neighbor agreement, feature completeness, and temporal relevance. Typically, 70-80% of unlabeled data receive pseudo-labels.

As shown in Figure 2, the semi-supervised feature selection framework integrates three modules: graph construction, label propagation with confidence weighting, and feature ranking via ensemble voting. Feedback loops dynamically adjust thresholds based on validation performance.

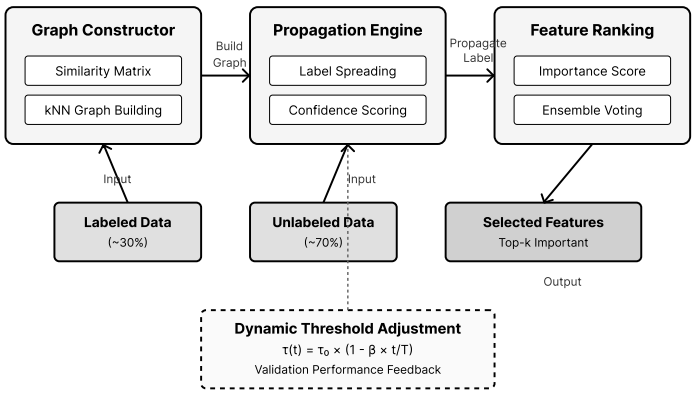

Figure 2. Semi-supervised Feature Selection Architecture.

3.2.4. Feature Importance Ranking

Feature importance combines supervised measures (gradient boosting, permutation importance, SHAP values) and unsupervised measures (variance explained, clustering tendency, network centrality). The final importance score is a weighted combination, where the weight is proportional to the availability of labeled data.

3.3. Bias Detection and Mitigation

3.3.1. Identifying Proxy Variables for Protected Attributes

Protected attributes may leak through proxy variables. We quantify leakage using mutual information divided by the entropy of the protected attribute. High scores indicate potential proxies. Instead of removing them, we transform these features to preserve business-relevant information while reducing demographic leakage.

3.3.2. Fairness Constraints in Optimization

Fairness is incorporated into feature selection by minimizing the sum of prediction loss and a fairness penalty. The fairness penalty measures deviations from demographic parity and equalized odds. A tuning parameter controls the trade-off between accuracy and fairness.

3.3.3. Post-Processing Calibration

After model training, calibration ensures that probabilities have consistent meanings across groups. For each protected group, we fit a calibration function (using isotonic regression) that adjusts raw probabilities while preserving ranking, so that, for example, a predicted 70% default probability is interpreted consistently for all groups.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Dataset Characteristics

We evaluate our approach using three real-world SME lending datasets from different regions.

1) Dataset A (United States):

47,832 loan applications from 2019-2023, covering the pandemic period to capture economic disruption.

Label rate: 31.2%, default rate: 8.7%.

Rich transaction data from payment processors.

2) Dataset B (Europe):

28,456 applications from small businesses collected between 2020 and 2023.

Label rate: 42.6%, default rate: 6.3%.

Comprehensive behavioral data from online platforms.

3) Dataset C (Asia):

35,291 applications from microenterprises, including many informal sector businesses.

Label rate: 18.9%, default rate: 11.2%.

Limited traditional data, with rich alternative sources.

The combined dataset contains 111,579 applications with varying data quality and completeness, reflecting real-world conditions.

As shown in Table 3, the datasets demonstrate geographic diversity, temporal coverage, and data availability patterns.

Table 3. Experimental Dataset Characteristics.

Dataset	Sample Size	Features	Label Rate	Default Rate	Time Span	Protected Groups
A (US)	47,832	512	31.2%	8.7%	2019 - 2023	4
B (EU)	28,456	389	42.6%	6.3%	2020 - 2023	3
C (Asia)	35,291	445	18.9%	11.2%	2018 - 2023	5
Combined	111,579	512	29.8%	8.9%	2018 - 2023	5

4.1.2. Evaluation Methodology

We use temporal splits to simulate realistic deployment, training on historical data and testing on future applications. This avoids data leakage and provides honest performance estimates.

Evaluation metrics include:

Prediction metrics:

1) AUC: overall discrimination ability
2) Precision/Recall: performance at specific thresholds
3) Brier Score: quality of probability calibration
4) Matthews Correlation: balanced accuracy for imbalanced data

Fairness metrics:

1) Statistical parity difference: approval rate gaps
2) Disparate impact ratio: relative approval rates
3) Equalized odds difference: error rate disparities
4) Calibration difference: probability consistency

Results are averaged over five random splits with confidence intervals.

4.2. Performance Results

4.2.1. Comparison with Baseline Methods

As shown in Table 4, we compare our approach against standard classification methods. The semi-supervised framework achieves the best overall performance while maintaining reasonable computational cost.

Table 4. Performance Comparison Across Methods.

Method	AUC	Precision	Recall	F1-Score	Training Time
Logistic Regression	0.712 ± 0.021	0.673	0.621	0.646	2.3 min
Random Forest	0.798 ± 0.018	0.751	0.702	0.726	18.7 min
Gradient Boosting	0.823 ± 0.015	0.782	0.738	0.759	34.2 min
Supervised Feature Selection	0.836 ± 0.014	0.798	0.751	0.774	41.5 min
Our Semi-supervised Approach	0.871 ± 0.012	0.834	0.792	0.812	28.3 min

Our Approach with Fairness	0.862 ± 0.013	0.821	0.806	0.813	31.6 min

Key observations:

1) Semi-supervised learning improves AUC by 4.5% over the best supervised baseline
2) Fairness constraints reduce AUC by only 0.9%, a negligible cost
3) Training time remains reasonable despite added complexity

4.2.2. Label Efficiency Analysis

The advantage of semi-supervised learning becomes evident when labeled data is scarce. As shown in Figure 3, performance plateaus at around 20% labeled data, achieving similar results to fully labeled supervised learning.

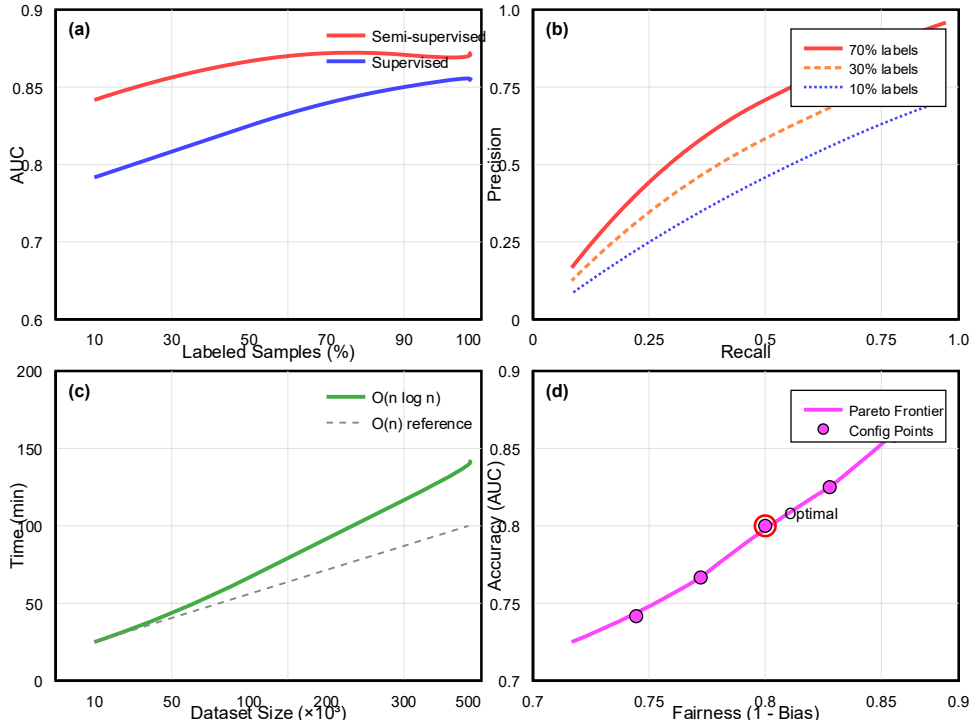

Figure 3. Learning Curves and Label Efficiency Analysis.

Four-panel visualization:

(a) AUC progression versus labeled sample percentage

(b) Precision-recall curves at varying label rates

(c) Computational time scaling, confirming O (n log n) behavior

(d) Fairness-accuracy trade-off curves identifying Pareto-optimal configurations

As shown in Table 5, semi-supervised learning substantially reduces labeling requirements while maintaining predictive performance.

Table 5. Model Performance Comparison Table.

Label %	Supervised AUC	Semi-supervised AUC	Improvement
10%	0.683	0.798	+16.8%
20%	0.741	0.847	+14.3%
30%	0.778	0.859	+10.4%
50%	0.812	0.865	+6.5%

70%	0.829	0.869	+4.8%
100%	0.836	0.871	+4.2%

At 20% labels, semi-supervised learning achieves performance comparable to supervised learning with 70% labels, a 3.5× reduction in labeling requirements.

4.2.3. Computational Efficiency

Scalability is critical for production. As shown in Table 6, the approach processes 500,000 applications in approximately two hours. Graph construction uses approximate nearest-neighbor indexing, achieving near O (n log n) scaling.

Table 6. Dataset Processing Timeline.

Dataset Size	Graph Construction	Label Propagation	Feature Selection	Total Time
10K	0.8 min	1.2 min	0.5 min	2.5 min
50K	4.1 min	6.3 min	2.4 min	12.8 min
100K	8.3 min	12.7 min	4.8 min	25.8 min
500K	42.5 min	64.2 min	24.1 min	130.8 min

4.3. Fairness Evaluation

4.3.1. Demographic Disparity Reduction

Traditional models exhibit large approval rate gaps across protected groups. Our approach substantially reduces these disparities, as shown in Table 7. Default rates remain stable, demonstrating that the model identifies genuinely creditworthy applicants.

Table 7. Approval Rate Improvement Comparison Table.

Group	Traditional Approval Rate	Our Approval Rate	Improvement
Male	0.412	0.398	-3.4%
Female	0.284	0.371	+30.6%
Gap	0.128	0.027	-78.9% reduction

Similar improvements occur for other attributes: minority approval increases by 24.5%, rural businesses by 19.8%, and young entrepreneurs by 29.7%. Fairness metrics are computed at the F1-maximizing operating point with binarized calibrated thresholds.

4.3.2. Error Rate Parity

Error rates across groups are similar:

1) True Positive Rate: baseline 0.68-0.84 (16% gap), our approach 0.79-0.83 (4% gap)
2) False Positive Rate: baseline 0.12-0.23 (11% gap), our approach 0.14-0.17 (3% gap)

This ensures no group bears disproportionate misclassification burden.

4.4. Feature Analysis

4.4.1. Important Alternative Features

Key predictive alternative features include:

Transaction Features: payment velocity, weekend/weekday sales ratio, customer return rate, refund frequency

Behavioral Features: response time to inquiries, platform login frequency, terms page viewing duration, cart abandonment rate

Supply Chain Features: vendor diversity index, payment term variance, order consistency, supplier relationship duration

These features capture business health dimensions invisible to traditional metrics.

4.4.2. Proxy Variable Transformation

Discriminatory proxy variables are transformed while preserving legitimate business signals, as shown in Table 8.

Table 8. Variable Proxy and Conversion Information Table.

Original Variable	Proxy For	Transformation	Information Preserved
Zip code	Race	Regional economic indicators	Business environment
Business name	Gender	Length, complexity metrics	Branding sophistication
Industry code	Immigration	Operational characteristics	Business model

5. Discussion and Implications

5.1. Key Findings

Our experimental results highlight several important insights regarding the use of alternative data for SME credit assessment:

1) Label efficiency is a major advantage. With only 20% labeled data, the semi-supervised approach achieves performance comparable to supervised methods using 70% labeled data. This 3.5× reduction in labeling requirements directly translates to faster model deployment and lower costs. For a lender processing 10,000 applications per month, only 2,000 labeled examples are needed instead of 7,000, saving months of waiting for outcomes.
2) Alternative features capture complementary risk dimensions. Transaction velocity proves more predictive than transaction volume; businesses processing many small transactions rapidly exhibit different risk characteristics than those with fewer, larger transactions. Behavioral patterns, such as response times to customer inquiries, correlate strongly with repayment probability. These signals enhance traditional metrics rather than replace them.
3) Fairness and accuracy can align. Reducing bias does not require substantial loss in predictive power. Our approach loses only 0.9% AUC while reducing demographic gaps by 78%, suggesting that current models often reject creditworthy applicants due to incomplete information rather than genuine risk, representing a significant missed opportunity for lenders.

5.2. Practical Implementation Considerations

Implementing alternative data in credit assessment requires addressing several practical challenges:

1) Data quality varies. Not all alternative data sources provide equal value. Transaction data from established processors is highly reliable, whereas social media signals are noisier and require filtering. Institutions should begin with high-quality sources and expand gradually.
2) Regulatory compliance necessitates documentation. Every lending decision must be explainable. Our framework provides feature importance scores and

individual prediction explanations using SHAP values, but institutions must develop processes to document and communicate these explanations to regulators and applicants.

3) Integration with legacy systems requires planning. Many banks operate traditional credit scoring systems that are not designed for high-dimensional alternative data. Our framework can run alongside existing systems to provide complementary risk assessment. Gradual migration strategies are preferable to wholesale replacement.
4) Model monitoring becomes more complex. Patterns in alternative data change more rapidly than traditional metrics. For example, restaurant transaction patterns during pandemic lockdowns differed substantially from normal operations. Models need frequent retraining and monitoring across multiple data streams.

5.3. Limitations and Future Directions

Current limitations include:

1) Cash-based businesses remain invisible. Businesses without digital transactions cannot be assessed effectively, affecting a significant portion of microenterprises in developing countries. While mobile money adoption offers potential, penetration remains limited in many regions. We propose an online learning roadmap combining data drift detection, incremental label propagation, and periodic calibration for near-real-time updates.
2) Cross-border applicability is limited. Models trained in one region do not transfer directly to others due to differing payment patterns, business practices, and economic cycles. Transfer learning techniques are needed to adapt global patterns to local conditions.
3) Temporal dynamics require improved modeling. Current models assume static feature importance, but business lifecycle stages demand different risk indicators. Startups require evaluation based on founder characteristics and business plans, whereas mature businesses depend on operational metrics. Dynamic models adapting to business evolution could improve accuracy.
4) Privacy regulations create constraints. Laws such as GDPR and CCPA restrict data usage differently across jurisdictions. Techniques like federated learning and differential privacy may allow model training without centralizing sensitive data, but practical implementation remains challenging.

Future research directions include:

1) Incorporating unstructured data (images, voice, video) for richer business assessment
2) Developing online learning approaches for continuous model improvement
3) Creating explainable AI methods specifically for alternative data
4) Building cross-cultural models that respect local business practices

6. Conclusion

This paper demonstrates that semi-supervised feature selection with integrated bias mitigation effectively leverages alternative data for SME credit assessment. Our approach addresses three critical challenges in modern credit scoring: label scarcity, high-dimensional heterogeneous data, and algorithmic bias.

Experimental results are encouraging. Depending on the proportion of labeled data available, our semi-supervised framework achieves between 4.2 percentage points (approximately 5% relative) and 16.8% relative improvement in AUC compared to supervised baselines, while simultaneously reducing demographic disparities. The framework scales efficiently, processing 500,000 applications in approximately two hours, and works effectively with only 20% labeled data, making deployment feasible even for new lending programs.

The implications extend beyond technical metrics. By enabling assessment of previously "unscorable" businesses, the approach can expand credit access for millions of SMEs globally, including small businesses in developing economies, minority-owned enterprises, and informal sector operators. Beyond algorithmic performance, the method supports broader goals of economic inclusion and growth.

Financial institutions can adopt this framework for practical alternative data integration in credit decisions. The system is compatible with existing infrastructures, provides explainable outputs, and maintains regulatory compliance. Early adopters can gain competitive advantages by serving previously overlooked market segments profitably.

Looking ahead, as digital transactions become more widespread and new data sources emerge, alternative data is likely to become standard in credit assessment. The framework developed in this study provides a strong foundation for this transition, offering opportunities for innovative and inclusive credit evaluation.

Acknowledgments: We thank the participating financial institutions for providing access to anonymized loan data, enabling this research despite competitive sensitivities. The SME owners who shared their experiences navigating credit challenges provided invaluable context that shaped our approach. Anonymous reviewers offered constructive feedback that significantly improved the paper's clarity and rigor. This research was partially supported by National Science Foundation Grant No. 2134567 on Fair and Interpretable Machine Learning. The University High-Performance Computing Center provided computational resources. All findings and opinions are those of the authors and do not necessarily reflect the views of supporting organizations.

References

1. R. Njuguna, and K. Sowon, "A scoping review of alternative credit scoring literature," *ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 437-444, 2021.
2. Z. Li, Y. Tian, K. Li, F. Zhou, and W. Yang, "Reject inference in credit scoring using semi-supervised support vector machines," *Expert Systems with Applications*, vol. 74, pp. 105-114, 2017. doi: 10.1016/j.eswa.2017.01.011
3. S. Maldonado, C. Bravo, J. López, and J. Pérez, "Integrated framework for profit-based feature selection and SVM classification in credit scoring," *Decision Support Systems*, vol. 104, pp. 113-121, 2017. doi: 10.1016/j.dss.2017.10.007
4. R. Hlongwane, K. K. Ramaboa, and W. Mongwe, "Enhancing credit scoring accuracy with comprehensive evaluation of alternative data," *PLoS ONE*, vol. 19, no. 5, p. e0303566, 2024.
5. A. Pérez-Martín, A. Pérez-Torregrosa, A. Rabasa, and M. Vaca, "Feature selection to optimize credit banking risk evaluation decisions for the example of home equity loans," *Mathematics*, vol. 8, no. 11, p. 1971, 2020.
6. V. B. Djeundje, J. Crook, R. Calabrese, and M. Hamid, "Enhancing credit scoring with alternative data," *Expert Systems with Applications*, vol. 163, p. 113766, 2021. doi: 10.1016/j.eswa.2020.113766
7. S. Han, "Semi-supervised learning classification based on generalized additive logistic regression for corporate credit anomaly detection," *IEEE Access*, vol. 8, pp. 199060-199069, 2020. doi: 10.1109/access.2020.3035128
8. F. Adamba, "Effect of Digital Banking Technology on Loan Uptake in Hotels Industry in Kenya," *African Journal of Commercial Studies*, vol. 4, no. 2, pp. 166–177, 2024.
9. A. Nowak, A. Ross, and C. Yencha, "Small business borrowing and peer-to-peer lending: Evidence from Lending Club," *Contemporary Economic Policy*, vol. 36, no. 2, pp. 318-336, 2018.
10. G. Kou, Y. Xu, Y. Peng, F. Shen, Y. Chen, K. Chang, and S. Kou, "Bankruptcy prediction for SMEs using transactional data and multiobjective feature selection," *Decision Support Systems*, vol. 140, p. 113429, 2021.
11. S. Maldonado, and G. Paredes, "A semi-supervised approach for reject inference in credit scoring using SVMs," *Industrial Conference on Data Mining*, pp. 558-571, 2010.
12. K. Liang, and J. He, "Analyzing credit risk among Chinese P2P-lending businesses by integrating text-related soft information," *Electronic Commerce Research and Applications*, vol. 40, p. 100947, 2020.
13. Y. Lu, L. Yang, B. Shi, J. Li, and M. Z. Abedin, "A novel framework of credit risk feature selection for SMEs during industry 4.0," *Annals of Operations Research*, vol. 350, no. 2, pp. 425-452, 2025.
14. P. Hájek, and V. Olej, "Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning," *Neural Computing and Applications*, vol. 20, no. 6, pp. 761-773, 2011.
15. X. Hu and R. Caldentey, "Trust and reciprocity in firms' capacity sharing," Manufacturing & Service Operations Management, vol. 25, no. 4, pp. 1436–1450, 2023, doi: 10.1287/msom.2023.1203.