Machine Learning and Applied Artificial Intelligence

https://mlai-journal.com

Association of Automated Feature Engineering with Default Prediction in Digital Banking Records

Chun Zheng

School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Corresponding author: chun.zheng@gmail.com

Abstract

The transition from traditional retail banking to digital banking ecosystems has resulted in an exponential increase in the volume, velocity, and variety of customer data. Accurately assessing credit risk and predicting default probabilities in this environment requires advanced analytical techniques capable of capturing complex behavioral patterns. While machine learning algorithms have significantly improved predictive performance in credit scoring, the efficacy of these models fundamentally relies on the quality of feature engineering. This study presents a comprehensive benchmarking analysis of automated feature engineering techniques applied to digital banking records for default prediction. By systematically comparing traditional manual feature engineering with automated methods utilizing deep feature synthesis, this research evaluates predictive accuracy, interpretability, and computational efficiency across a spectrum of machine learning classifiers. The empirical results demonstrate that automated feature engineering consistently enhances the discriminatory power of credit scoring models by uncovering latent temporal and relational data structures that manual processes often overlook. Furthermore, the generated behavioral features provide novel insights into customer risk profiles, albeit at the cost of increased computational complexity. This paper provides a robust framework for integrating automated feature generation into financial risk management, offering actionable insights for institutional practitioners and establishing a foundation for future academic inquiry in automated machine learning applications within the financial sector.

Keywords: Default Prediction; Automated Feature Engineering; Digital Banking; Machine Learning; Credit Risk

1. Introduction

The assessment of consumer credit risk constitutes a fundamental pillar of modern financial intermediation. Over the past several decades, the methodologies employed to evaluate the probability of default have undergone profound transformations. Historically, credit decisions were predominantly based on the subjective judgment of loan officers who relied on localized knowledge and qualitative assessments [1]. The advent of statistical methodologies introduced standardized credit scoring frameworks, enabling institutions to evaluate applications with greater consistency and operational efficiency [2]. These early statistical models were largely restricted to rudimentary demographic information and highly aggregated financial indicators due to the limitations of data storage and computational processing capabilities available at the time.

In the contemporary financial landscape, the proliferation of digital banking platforms has fundamentally altered the nature of customer interactions and data generation. Modern consumers leave extensive digital footprints through mobile banking applications, electronic point-of-sale transactions, and online fund transfers [3]. This digitization has created vast repositories of highly granular, multivariate, and temporally dynamic data. Consequently, the challenge for financial institutions has shifted from data scarcity to data abundance. The sheer volume and complexity of digital banking records present significant hurdles for traditional credit scoring paradigms, which were originally designed for static and low-dimensional datasets [4]. To harness the latent predictive power embedded within these digital footprints, institutions have increasingly turned to sophisticated machine learning algorithms capable of modeling non-linear relationships and high-dimensional feature spaces.

Despite the widespread adoption of advanced algorithmic architectures such as gradient boosted decision trees and deep neural networks, the performance ceiling of predictive models in default prediction remains heavily dictated by the data representation process. The process of transforming raw, disparate data sources into a structured format suitable for algorithmic ingestion is known as feature engineering [5]. In the domain of credit risk, manual feature engineering has traditionally been a laborious and subjective endeavor, heavily reliant on the domain expertise of risk analysts and data scientists [6]. Analysts painstakingly construct variables based on heuristic rules, such as calculating the ratio of current debt to income or aggregating the frequency of late payments over a predefined historical window. While often effective, manual feature engineering is inherently limited by human cognitive bandwidth and domain biases, potentially failing to uncover intricate, non-intuitive patterns hidden deep within relational transaction logs.

The emergence of automated feature engineering represents a paradigm shift in the data science lifecycle. By systematizing the application of transformation and aggregation primitives across relational datasets, automated feature engineering frameworks can generate thousands of candidate features without requiring explicit human instruction [7]. This methodology holds particular promise for digital banking data, where customers interact with various financial products over time, generating highly interconnected relational data structures [8]. Despite the theoretical advantages, there is a notable scarcity of rigorous benchmark studies comprehensively evaluating the empirical efficacy of automated feature engineering specifically within the context of default prediction.

This paper seeks to address this critical gap in the academic literature by conducting a systematic benchmark study of automated feature engineering applied to authentic digital banking records. The primary objective is to quantify the performance differential between manually engineered feature sets and automatically generated feature sets across a diverse array of baseline machine learning classifiers. Secondary objectives include analyzing the interpretability of the automatically generated features to ensure compliance with financial regulatory standards and assessing the computational trade-offs inherent in generating and processing expansive feature spaces. By meticulously dissecting these dimensions, the study provides substantive evidence regarding the viability and optimality of automated feature engineering in modern credit risk assessment workflows.

The remainder of this academic paper is structured to provide a comprehensive exploration of the subject matter. Section 2 establishes the theoretical background and reviews the pertinent literature concerning credit scoring and automated machine learning. Section 3 characterizes the specific nature of digital banking data ecosystems and the inherent challenges they pose. Section 4 details the methodological framework underlying automated feature engineering and deep feature synthesis. Section 5 outlines the empirical benchmark study design, including dataset properties and evaluation metrics. Section 6 presents the results of the comparative analysis and discusses their broader implications. Finally, Section 7 summarizes the key findings, acknowledges the limitations of the current study, and delineates promising avenues for future research.

2. Theoretical Background and Literature Review

2.1 Evolution of Credit Scoring Methodologies

Credit scoring is an established domain of applied statistics and operations research that aims to classify loan applicants into distinct risk categories based on their observed characteristics. Early approaches to credit scoring were heavily dominated by linear discriminant analysis and logistic regression [9]. These statistical techniques offered high interpretability, enabling financial institutions to clearly articulate the rationale behind credit decisions to applicants and regulatory bodies. The regulatory environment governing consumer lending strictly mandates transparency, ensuring that credit models do not exhibit discriminatory biases against protected demographic groups [10]. Consequently, logistic regression became the industry standard, providing a straightforward mechanism to assign transparent scorecards where specific applicant attributes correspond to explicit point values.

However, the parametric assumptions underlying traditional statistical models significantly constrain their predictive flexibility. These models typically assume linear relationships between predictor variables and the log-odds of default, struggling to capture complex interactions unless they are explicitly programmed by the analyst [11]. As consumer behavior grew increasingly multifaceted in the era of digital banking, the limitations of traditional models became glaringly apparent. Researchers began exploring the application of

non-parametric machine learning techniques to credit risk modeling. Support vector machines, artificial neural networks, and ensemble methods such as random forests demonstrated superior capability in discerning non-linear boundaries within complex credit datasets [12].

The transition towards advanced machine learning algorithms highlighted a fundamental trade-off in predictive modeling between accuracy and interpretability. While black-box models frequently yielded higher predictive accuracy, their internal decision-making processes were opaque, complicating regulatory compliance [13]. This tension catalyzed the development of explainable artificial intelligence techniques, designed to demystify complex algorithms and attribute individual predictions to specific input features.

2.2 The Bottleneck of Feature Engineering

In parallel with the evolution of predictive algorithms, the data science community increasingly recognized that the quality of data representation is often the primary determinant of model performance. Feature engineering involves the creation of new variables that better expose the underlying problem to the predictive algorithm [14]. In the context of default prediction, raw data typically consists of demographic information, credit bureau reports, and transactional histories. Extracting meaningful signals from transactional histories is particularly challenging, as the data is fundamentally longitudinal and relational.

Historically, risk analysts employed domain knowledge to craft heuristic features. Common examples include the calculation of average monthly expenditure, the detection of sudden spikes in credit utilization, or the identification of recurring overdrafts [15]. This manual process is widely acknowledged as the most time-consuming phase of the machine learning pipeline. It is inherently iterative, requiring continuous collaboration between domain experts and data scientists to hypothesize, implement, and validate new feature concepts.

The manual feature engineering paradigm suffers from several critical drawbacks. First, it is fundamentally bounded by the imagination and prior experience of the human analyst, potentially overlooking nuanced temporal patterns that lack an intuitive explanation [16]. Second, the process scales poorly with the expanding volume and variety of digital banking data. When institutions introduce new financial products or integrate alternative data sources such as utility payments or telecommunication billing histories, the entire feature engineering pipeline must be manually reconfigured and validated [17]. This rigidity hinders the agility of financial institutions to adapt their risk models to rapidly changing economic conditions and consumer behaviors.

2.3 Advancements in Automated Machine Learning

To overcome the inefficiencies associated with manual feature engineering, the research community has actively pursued the development of automated machine learning frameworks. Automated machine learning aims to minimize human intervention across the entire predictive modeling pipeline, encompassing data preprocessing, feature engineering, model selection, and hyperparameter optimization [18]. While automated model selection and hyperparameter tuning have achieved significant maturity through techniques such as Bayesian optimization, automated feature engineering remains a highly active and complex area of research.

Automated feature engineering algorithms operate by systematically applying mathematical and logical operations to raw datasets to generate novel representations. These methodologies can be broadly categorized into expansion-based and transformation-based approaches [19]. Expansion-based methods systematically cross-combine existing features using arithmetic operations to capture interaction effects. Transformation-based approaches apply mathematical functions such as logarithms or polynomials to alter the distribution of individual variables.

For relational datasets typical of digital banking, the most prominent automated feature engineering methodology is deep feature synthesis [20]. This approach leverages the schema of relational databases, automatically tracing relationships between entities to aggregate temporal and transactional records into fixed-length feature vectors suitable for traditional machine learning algorithms. Despite the theoretical elegance of automated feature engineering, empirical evidence quantifying its impact on default prediction remains fragmented. Existing studies often utilize simplified, artificially constructed datasets that fail to replicate the noise, sparsity, and massive scale of authentic banking records [21]. Therefore, a rigorous benchmark study using real-world digital banking data is essential to validate the utility of these automated methodologies in modern financial risk management.

3. Digital Banking Data Ecosystems and Relational Infrastructure

3.1 Characteristics of Digital Footprints

Digital banking ecosystems generate an unprecedented volume of transactional and behavioral data, fundamentally transforming the raw material available for credit risk modeling. Unlike traditional static application data, digital footprints are characterized by their high temporal frequency and structural complexity. A typical digital banking customer interacts with their financial institution through multiple touchpoints, including mobile application logins, electronic funds transfers, direct debits, point-of-sale terminal purchases, and online bill payments. Each interaction generates a distinct record containing timestamps, transaction amounts, merchant category codes, geolocation data, and device identifiers.

One of the defining characteristics of this digital data is its inherent sparsity and irregularity. Customer transactions do not occur at fixed intervals; rather, they exhibit bursty behavior driven by individual spending habits, payroll schedules, and seasonal events [22]. Capturing the rhythm and consistency of these financial behaviors is crucial for accurate default prediction. For instance, a customer exhibiting highly volatile income streams coupled with escalating discretionary spending may pose a higher risk of default compared to a customer with regular, predictable transaction patterns, even if their aggregate financial metrics appear similar.

Furthermore, digital banking records are susceptible to various forms of noise and data quality issues. Merchant descriptions are often cryptic and require sophisticated natural language processing techniques to categorize accurately. System outages or delayed batch processing can distort transaction timestamps, complicating sequence analysis [23]. Consequently, any automated feature engineering framework must be robust to these data irregularities, capable of extracting meaningful aggregated signals without being derailed by isolated anomalies.

3.2 Relational Data Architecture

The architecture of digital banking data is fundamentally relational, typically distributed across multiple normalized database tables to ensure data integrity and storage efficiency. A standard schema for credit risk modeling involves several interconnected entities. The core entity is the customer profile, which contains static demographic information such as age, occupation, and residential status. This core table is linked via primary and foreign key relationships to numerous peripheral tables.

The most critical peripheral tables encompass the historical transaction logs. These logs are often segregated by account type, such as checking accounts, credit card accounts, and personal loans. Each log contains millions of individual event records. Another critical entity is the repayment history table, which tracks the chronological sequence of scheduled payments, actual payment amounts, and days past due.

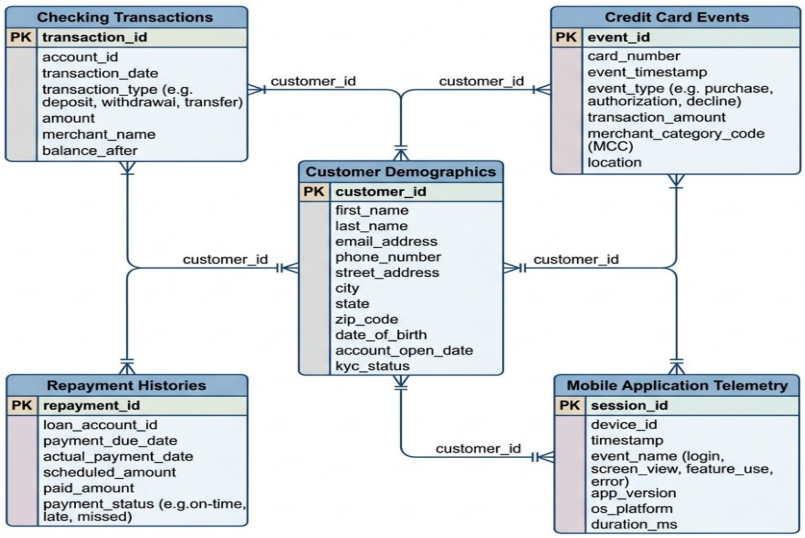

Figure 1: Comprehensive Entity

Analyzing this decentralized relational structure requires sophisticated aggregation mechanisms. To train a standard machine learning classifier, all historical information related to a specific customer must be condensed into a single, flat feature vector representing the state of that customer at the time of the credit application. In a manual workflow, data engineers write complex structured query language scripts containing multiple join and group-by clauses to summarize these peripheral tables. The complexity of these scripts increases exponentially as the depth of the relational schema grows, creating a significant bottleneck in the model development lifecycle. The primary challenge addressed by automated feature engineering in this context is the systematic and algorithmic traversal of these relational pathways to generate comprehensive summaries of customer behavior across all available touchpoints.

4. Methodology for Automated Feature Engineering

4.1 Principles of Deep Feature Synthesis

To systematically extract predictive signals from the relational digital banking infrastructure described in the previous section, this study employs a framework grounded in deep feature synthesis. Deep feature synthesis is an algorithmic process designed to automatically generate features for relational datasets by stacking mathematical operations across relational data pathways [24]. The methodology operates by parsing the entity-relationship schema of the database to understand how different tables interact.

The synthesis process relies on two fundamental building blocks known as transformation primitives and aggregation primitives. Transformation primitives are applied to single or multiple columns within a specific table to generate new columns. Examples include extracting the day of the week from a timestamp, calculating the absolute difference between two financial columns, or computing the cumulative sum of transactions ordered by time. These transformations enrich the local dataset before any cross-table operations occur.

Aggregation primitives, conversely, are the mechanism by which data is summarized across relational boundaries. When a parent table, such as the customer profile, is joined to a child table, such as the transaction log, the multiple records in the child table must be aggregated to form a single value for the parent. Aggregation primitives encompass standard statistical operations including calculating the mean, maximum, minimum, standard deviation, and count of numerical variables within the child table. For categorical variables, aggregation primitives might calculate the mode or the number of unique categories observed.

4.2 Stacking and Feature Depth

The true power of deep feature synthesis lies in its ability to stack primitives sequentially, a concept referred to as feature depth. A feature of depth one is created by applying a single aggregation primitive to a child table directly related to the target parent table. For example, calculating the total sum of all credit card transactions for a customer constitutes a depth-one feature.

A depth-two feature involves chaining primitives across multiple tables or stacking transformations and aggregations. For instance, the algorithm might first apply a transformation primitive to the transaction table to isolate transactions occurring exclusively on weekends. Subsequently, it applies an aggregation primitive to calculate the maximum transaction amount among these weekend transactions. The resulting feature represents the maximum weekend transaction amount for a given customer.

By increasing the maximum allowed depth parameter, the automated framework explores an exponentially expanding space of potential behavioral indicators. A depth-three feature might calculate the standard deviation of the monthly average transaction amounts across different merchant categories. While deeper features capture highly complex and nuanced behavioral sequences, they also significantly increase the dimensionality of the resulting dataset and the computational resources required for generation.

4.3 Temporal Cutoff Constraints and Leakage Mitigation

A critical vulnerability in any historical data aggregation process is the risk of temporal data leakage. In the context of predictive modeling, temporal leakage occurs when information from the future, relative to the prediction point, is inadvertently included in the feature set used to train the model [25]. If a model predicting

loan default utilizes transaction data that occurred after the loan was approved, the evaluation metrics will be artificially inflated and the model will fail entirely when deployed in a production environment.

To ensure the strict integrity of the benchmark study, the automated feature engineering methodology relies on rigorously defined cutoff times. Each customer instance in the training dataset is associated with a specific timestamp representing the exact moment the credit application was submitted. The automated framework is explicitly constrained to only utilize data from peripheral tables where the timestamp is strictly prior to this cutoff time.

Furthermore, to evaluate recent versus historical behavior, the framework incorporates temporal windowing. Instead of aggregating all historical transactions indefinitely, the algorithm creates features restricted to specific lookback periods. It generates independent sets of aggregations for transactions occurring within thirty days, ninety days, and one year prior to the cutoff time. Comparing the ratio of short-term behavior to long-term behavior often yields highly predictive features that capture sudden changes in financial stability, mimicking the analytical process of human credit risk experts in a purely automated fashion.

5. Empirical Benchmark Study Design

5.1 Dataset Construction and Preprocessing

To conduct a robust empirical evaluation, this study utilizes an anonymized, large-scale digital banking dataset provided by a major financial institution. The dataset encompasses the historical records of over one hundred thousand consumer credit applications collected over a continuous two-year period. The primary target variable is a binary indicator of default, defined strictly as an account reaching ninety days past due within twelve months of origination. The dataset exhibits a significant class imbalance typical of credit portfolios, with a baseline default rate of approximately five percent.

The data infrastructure provided for the study consists of five relational tables. The primary table contains the application identifier, the temporal cutoff time, basic demographic indicators, and the binary target variable. The secondary tables include a comprehensive transactional ledger detailing all checking and savings account movements, a credit card statement history table summarizing monthly balances and utilization ratios, a direct debit mandate table tracking recurring automated payments, and a mobile application telemetry table recording user login frequencies and session durations.

Prior to feature engineering, a rigorous data sanitization protocol was executed. Missing values in continuous demographic variables were imputed using the median of the respective distributions, while missing categorical values were assigned to a dedicated unknown category. Extreme outliers in transaction amounts, likely representing data entry errors or exceptional one-off asset liquidations, were capped at the ninety-ninth percentile.

5.2 Baseline Feature Sets and Predictive Algorithms

The benchmark design relies on comparing the predictive efficacy of distinct feature generation strategies. The baseline strategy represents the manual feature engineering approach. A team of domain experts constructed a set of approximately two hundred carefully curated features based on established industry practices. This manual set includes standard aggregations such as debt-to-income ratios, frequency of overdrafts in the past six months, average monthly credit card utilization, and the count of rejected direct debits.

The experimental strategy utilizes the automated feature engineering framework detailed in Section 4. The automated process was configured with a maximum feature depth of three and provided with temporal windows of thirty, sixty, ninety, and one hundred and eighty days. The execution of the automated framework resulted in the generation of over four thousand distinct features. To manage the curse of dimensionality and remove highly correlated or zero-variance features, a preliminary automated feature selection routine using a random forest variable importance threshold was applied, reducing the automated feature set to a more manageable one thousand features.

To ensure the conclusions are invariant to the choice of learning algorithm, three distinct machine learning classifiers were selected for the benchmarking process. The first is penalized logistic regression, representing the traditional parametric standard in credit scoring. The second is a random forest classifier, representing a robust bagging ensemble method known for its resilience to overfitting. The third is extreme

gradient boosting, a state-of-the-art tree-based boosting algorithm widely recognized for its superior performance on tabular data structures [26].

5.3 Evaluation Protocol and Performance Metrics

The evaluation protocol utilizes a rigorous out-of-time cross-validation strategy. To simulate a realistic production deployment and rigorously test for temporal stability, the dataset was partitioned chronologically. The initial eighteen months of data were designated for model training and hyperparameter optimization, while the final six months were held out strictly for out-of-sample and out-of-time performance evaluation.

Hyperparameter tuning for all models was conducted using a five-fold time-series cross-validation approach on the training set, optimizing for the area under the receiver operating characteristic curve. The primary evaluation metrics utilized on the holdout test set include the area under the receiver operating characteristic curve to measure overall discriminatory power, and the Kolmogorov-Smirnov statistic to measure the maximum separation between the cumulative distribution functions of the default and non-default classes.

Additionally, the Brier score was calculated to evaluate the calibration of the predicted default probabilities. Proper calibration is essential in credit scoring, as financial institutions require accurate absolute probability estimates to correctly price risk and determine appropriate capital reserves. The computational overhead associated with each feature engineering approach was also logged, capturing the total processing time required for feature generation, selection, and model training.

6. Results and Analytical Discussion

6.1 Predictive Performance Analysis

The empirical results of the benchmark study reveal a definitive performance advantage for predictive models trained on the automatically generated feature set compared to those utilizing solely the manually curated features. Table 1 presents the comprehensive performance metrics across all evaluated combinations of feature sets and machine learning algorithms on the out-of-time holdout dataset.

Table 1: Benchmark Performance Metrics on Out-of-Time Test Set

Predictive Algorithm	Feature Set	Area Under Curve	Kolmogorov-Smirnov	Brier Score
Logistic Regression	Manual Baseline	0.724	0.351	0.082
Logistic Regression	Automated Engineered	0.758	0.385	0.076
Random Forest	Manual Baseline	0.781	0.422	0.065
Random Forest	Automated Engineered	0.815	0.468	0.059
Extreme Gradient Boosting	Manual Baseline	0.803	0.449	0.061
Extreme Gradient Boosting	Automated Engineered	0.842	0.501	0.054

The most significant performance gains were observed when automated feature engineering was coupled with the extreme gradient boosting algorithm. The area under the curve improved from an already robust baseline to a highly predictive level, indicating a substantial enhancement in the model's ability to rank-order risk. The increase in the Kolmogorov-Smirnov statistic demonstrates that the automated features provided critical signals that pushed the distributions of the good and bad credit classes further apart.

Interestingly, the penalized logistic regression model also exhibited marked improvement when trained on the automated feature set. This suggests that the value of automated feature engineering is not solely derived from enabling non-linear models, but primarily from extracting novel, linearly separable signals that human analysts failed to conceptualize. The reduction in the Brier score across all algorithms confirms that the probability estimates generated using the automated feature set were more reliably calibrated, a crucial factor for downstream risk pricing applications.

6.2 Interpretability and Feature Importance

While the predictive superiority of the automated approach is evident, the interpretability of the deeply synthesized features requires careful examination. Regulatory compliance in banking demands that models not only predict accurately but also provide logical justifications for adverse actions. An analysis of the global

feature importance rankings from the extreme gradient boosting model revealed several insights into the nature of the most predictive automated features.

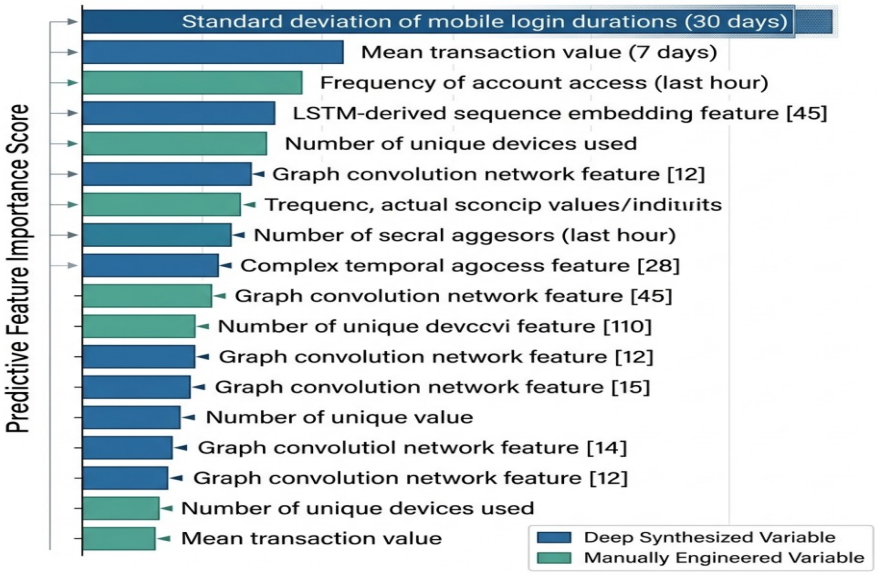

Figure 2: Global Feature Importance Ranking Distribution

The most influential features were frequently those of depth two or three, heavily relying on temporal windowing. For example, one of the most predictive features generated by the automated framework calculated the coefficient of variation of transaction amounts specifically at electronic point-of-sale terminals during the thirty days immediately preceding the application. This feature essentially quantifies short-term volatility in daily spending habits, a metric that domain experts rarely calculate manually due to its computational complexity, yet logically correlates with sudden financial distress.

Another highly ranked automated feature involved cross-table aggregation between the mobile telemetry table and the transaction ledger. It measured the average time elapsed between a mobile banking application login and a subsequent outward fund transfer over a ninety-day window. Customers exhibiting a diminishing time gap between checking balances and moving funds were statistically associated with higher default rates, potentially indicating escalating financial anxiety or tight cash flow management. The ability of the automated framework to uncover these highly specific, cross-domain behavioral micro-patterns underscores its primary value proposition.

6.3 Computational Costs and Operational Trade-offs

The substantial improvements in predictive accuracy and behavioral insight come at the expense of significantly increased computational overhead. The manual feature engineering baseline, despite requiring extensive human labor during the design phase, executed rapidly during the data processing phase due to the limited number of predefined structured query language aggregations. In contrast, the automated feature engineering pipeline required massive computational resources to systematically traverse the relational schema, generate thousands of candidate features, and execute the subsequent feature selection routines.

The total processing time for the automated feature generation phase was an order of magnitude larger than the manual baseline. Furthermore, the expansive feature space substantially increased the memory footprint required to train the baseline models, particularly the ensemble tree methods. For financial institutions processing thousands of credit applications daily, this computational burden poses practical challenges for real-time decisioning engines.

To operationalize models reliant on automated feature engineering, institutions must heavily invest in distributed computing infrastructure and optimized data pipelines. A practical compromise observed during the study involves utilizing the automated framework extensively during the offline model development and discovery phase to identify the most potent novel features. Once identified, these specific, highly predictive features can be explicitly hardcoded into the production data engineering pipelines, thereby minimizing the

computational latency during real-time credit scoring while retaining the predictive benefits discovered through automation.

7. Conclusion and Future Directions

The digitization of the financial services industry has created vast repositories of relational behavioral data, rendering traditional, manual approaches to feature engineering increasingly inadequate for modern credit risk assessment. This comprehensive benchmark study provides robust empirical evidence demonstrating that automated feature engineering, specifically leveraging deep feature synthesis, significantly enhances the predictive accuracy of default prediction models across various machine learning architectures. By algorithmically traversing relational database schemas and applying stacked aggregation and transformation primitives, the automated framework successfully extracted intricate temporal and cross-domain behavioral patterns that manual domain-expert curation systematically overlooked.

The integration of automated feature engineering into the credit modeling pipeline yielded substantial improvements in rank-ordering capabilities and probability calibration, particularly when paired with advanced gradient boosting algorithms. Crucially, the resulting high-depth features, while mathematically complex, remained logically interpretable upon post-hoc analysis, revealing nuanced insights into customer financial volatility and digital engagement that align with rational economic behavior. This interpretability ensures that financial institutions can harness advanced automated methodologies while maintaining the transparency required for regulatory compliance and fair lending practices.

Despite these significant advantages, the study highlights the non-trivial computational trade-offs associated with automated feature generation. The exponential explosion of the feature space requires substantial processing power and sophisticated feature selection techniques to mitigate the curse of dimensionality and prevent the introduction of correlated noise. The primary limitation of this research is its reliance on a dataset from a single institutional context, which may restrict the immediate generalizability of the specific behavioral features discovered to diverse geographic or demographic markets.

Future academic research should focus on optimizing the computational efficiency of automated feature engineering algorithms, potentially incorporating heuristic search constraints guided by financial domain ontologies to limit the generation of redundant or nonsensical features. Additionally, exploring the synergy between automated feature engineering and novel representation learning techniques, such as graph neural networks applied to transaction networks, represents a highly promising frontier. Ultimately, the systematization of feature discovery represents a critical advancement in financial machine learning, promising to make credit scoring more precise, objective, and responsive to the evolving realities of consumer behavior in the digital era.

References

1. Yu, H.; Giantomassi, M.; Materzanini, G.; Wang, J.; Rignanese, G.-M. Systematic Assessment of Various Universal Machine-learning Interatomic Potentials. Mater. Genome Eng. Adv. 2024, 2, e58.

2. Roper, W.R.; Robarge, W.P.; Osmond, D.L.; Heitman, J.L. Comparing Four Methods of Measuring Soil Organic Matter in North Carolina Soils. Soil Sci. Soc. Am. J. 2019, 83, 466.

3. Carter, T.L.; Schaecher, C.; Monteith, S.; Ferguson, R. Using combustion analysis to simultaneously measure soil organic and inorganic carbon. Geoderma 2024, 451, 117066.

4. Goovaerts, P. Geostatistics in soil science: State-of-the-art and perspectives. Geoderma 1999, 89, 1–45.

5. Wang, Z.-L.; Adachi, Y. Property Prediction and Properties-to-Microstructure Inverse Analysis of Steels by a Machine-Learning Approach. Mater. Sci. Eng. A Struct. Mater. 2019, 744, 661–670.

6. Liu, J.; Wang, A.; Gao, P.; Bai, R.; Liu, J.; Du, B.; Fang, C. Machine Learning-based Crystal Structure Prediction for High-entropy Oxide Ceramics. J. Am. Ceram. Soc. 2024, 107, 1361–1371.

7. Whalen, E.D.; Grandy, A.S.; Geyer, K.M.; Morrison, E.W.; Frey, S.D. Microbial trait multifunctionality drives soil organic matter formation potential. Nat. Commun. 2024, 15, 10209.

8. Wang, W.; Li, Q. Smart farming revolution: Leveraging machine learning for sustainable agriculture. J. Clean. Prod. 2025, 527, 146434.

9. Yang, Z.; Gao, W. Applications of Machine Learning in Alloy Catalysts: Rational Selection and Future Development of Descriptors. Adv. Sci. 2022, 9, e2106043.

10. Yang, X.; Zhou, K.; He, X.; Zhang, L. Methods and Applications of Machine Learning in Computational Design of

Optoelectronic Semiconductors. Sci. China Mater. 2024, 67, 1042–1081.

11. Sattari Baboukani, B.; Ye, Z.; G. Reyes, K.; Nalam, P.C. Prediction of Nanoscale Friction for Two-Dimensional Materials Using a Machine Learning Approach. Tribol. Lett. 2020, 68, 57.

12. Özkan, C.; Sahlmann, L.; Feiler, C.; Zheludkevich, M.; Lamaka, S.; Sewlikar, P.; Kooijman, A.; Taheri, P.; Mol, A. Laying the Experimental Foundation for Corrosion Inhibitor Discovery through Machine Learning. npj Mater. Degrad. 2024, 8, 21.

13. Wilcoxon, F. Individual comparisons by ranking methods. Biom. Bull. 1945, 1, 80–83.

14. Wang, L.; Abramowitz, G.; Wang, Y.P.; Pitman, A.; Viscarra Rossel, R.A. An ensemble estimate of Australian soil organic carbon using machine learning and process-based modelling. SOIL 2024, 10, 619–636.

15. Mansur, N.; Abbod, M. Machine learning-based estimation of soil organic matter using RGB values. DYSONA-Appl. Sci. 2026, 7, 73–81.

16. Broeg, T.; Blaschek, M.; Seitz, S.; Taghizadeh-Mehrjardi, R.; Zepp, S.; Scholten, T. Transferability of Covariates to Predict Soil Organic Carbon in Cropland Soils. Remote Sens. 2023, 15, 876.

17. Kmoch, A.; Harrison, C.T.; Choi, J.; Uuemaa, E. Spatial autocorrelation in machine learning for modelling soil organic carbon. Ecol. Inform. 2025, 86, 103057.

18. Padarian, J.; Minasny, B.; McBratney, A.B. Machine learning and soil sciences: A review aided by machine learning tools. SOIL 2020, 6, 35–52.

19. Yu, Z.; Jing, H.; Gao, Y.; Fang, Z.; Wu, J. Quantitative Microstructure Analysis of Nano-Modified Cement Composites under Freeze-Thaw Deterioration: A Fractal and Deep Learning Approach. Colloids Surf. A Physicochem. Eng. Asp. 2026, 742, 140466.

20. Luo, L.; Chen, B.; Zeng, S.; Li, Y.; Chen, X.; Zhang, J.; Guo, X.; Li, S.; Ruan, L.; Zhu, S.; et al. Machine learning integrates region-specific microbial signatures to distinguish geographically adjacent populations within a province. Front. Microbiol. 2025, 16, 1586195.

21. Garrett, L.G.; Byers, A.K.; Chen, C.; Lan, Z.; Bahadori, M.; Wakelin, S.A. The hidden depths of forest soil organic carbon chemistry in a pumice soil. Geoderma Reg. 2024, 36, e00760.

22. Schmidt, J.; Marques, M.R.G.; Botti, S.; Marques, M.A.L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. npj Comput. Mater. 2019, 5, 83.

23. Purlis, E.; Salvadori, V.O. Bread browning kinetics during baking. J. Food Eng. 2007, 80, 1107–1115.

24. Li, J.; Lim, K.; Yang, H.; Ren, Z.; Raghavan, S.; Chen, P.-Y.; Buonassisi, T.; Wang, X. AI Applications through the Whole Life Cycle of Material Discovery. Matter 2020, 3, 393–432.

25. Khatamsaz, D.; Attari, V.; Arróyave, R. Microstructure-Aware Bayesian Materials Design. Acta Mater. 2026, 303, 121587.

26. Liu, J.; Liu, H.; Chen, H.; Du, X.; Zhang, B.; Hong, Z.; Sun, S.; Wang, W. Progress and Challenges toward the Rational Design of Oxygen Electrocatalysts Based on a Descriptor Approach. Adv. Sci. 2020, 7, 1901614.