Development Economics of China
2025, VOL. 9, NO. 10, 67-70
DOI: 10.47297/wspdecWSP2515-797314.20250910

Ant Financial's Research on Risk Control Using Digital Footprints

Xinbo Yu

Department of Information Systems, College of Business, City University of Hong Kong, Hong Kong, 999077, China

ABSTRACT
In the era of deep integration between mobile internet and the digital economy, users' "digital footprints" — unstructured data encompassing online transaction preferences, social interactions, and other dimensions — are evolving into high-value information assets in financial risk management. Traditional credit evaluation systems, which primarily rely on static structured financial data, face inherent challenges such as information asymmetry and limited population coverage. This study examines how Ant Financial, a leading fintech enterprise, leverages massive user digital footprint data to construct innovative risk management models. The research develops a simulated dataset containing 50,000 samples that integrates traditional financial characteristics with digital footprint indicators reflecting user behavior. Two models are then designed for comparative analysis: a linear regression model based solely on traditional financial features serves as the benchmark, while a decision tree model incorporating digital footprints is implemented through feature correlation screening. Through comparative analysis of these models, this study demonstrates the pivotal role of digital footprints in expanding the reach of inclusive financial services and enhancing credit risk prediction accuracy. Research demonstrates that decision tree models incorporating digital footprints outperform traditional approaches in identifying potential default risks. The model structure also reveals soft information such as user behavior stability and consumption patterns, which holds significant value for credit assessment. This provides valuable insights for the digital transformation of traditional financial institutions. Furthermore, the study offers an analytical framework to understand the core risk control logic of fintech companies like Ant Financial.

KEYWORDS
Digital footprint; Risk management; Credit scoring; Ant Financial; Decision tree; Fintech

1 Introduction

The global financial landscape is being reshaped by the digital era with unprecedented depth and breadth. As social networks, e-commerce, and mobile payments continue to proliferate, human behavior increasingly exhibits digital characteristics, leaving massive "digital footprints" in cyberspace. These footprints include small-scale data trails like utility payments and QR code transactions, alongside large-scale data trajectories such as long-term social network profiles and consumption preferences. Together, they form a high-definition portrait that characterizes individual creditworthiness, habits, and behaviors. This data blue ocean containing the key to solving traditional risk control challenges holds immense significance for financial institutions [1-2].

For a long time, traditional credit evaluation systems such as the FICO score in the United States and China's credit reports have mainly relied on structured "hard information" like historical credit records, income certificates, and debt levels. However, this system has significant limitations. Borrowers may conceal their true financial status, making it difficult for financial institutions to assess their default risks, and the problem of information asymmetry remains severe. There are still billions of "thin-file individuals" globally, and the coverage of traditional credit reporting systems is limited. Due to the lack of effective credit history, they struggle to obtain fair financial services, which greatly hinders the realization of inclusive finance (Iyer, Khwaja, Luttmer, & Shue, 2016)[3].

Against this backdrop, the use of digital footprints as supplementary data to traditional credit information has emerged as a cutting-edge research focus in both academia and industry. These 'soft information' metrics offer a fresh perspective for risk assessment, as they can reflect an individual's repayment willingness, life stability, and even personality traits (Berg, Burg, Gombovic, & Puri, 2020)[4].

Leveraging the digital ecosystem comprising Alipay Sesame Credit, MYbank, and other platforms, Ant Financial—a global leader in fintech—has accumulated massive digital footprints of users. This study selects Ant Financial as a case study to investigate the practical effectiveness and underlying logic of utilizing digital footprints for risk management, conduct comparative experiments, and construct a simulated dataset to quantify the value of digital footprints in credit scoring. The paper is structured as follows: Part II reviews literature on traditional risk control and digital footprints, including data construction and model specifications; Part III outlines the research design and methodology; Part IV analyzes experimental results; Part V provides a summary and discusses the research significance, limitations, and future directions.

2 literature Review

2.1 Traditional Credit Risk Assessment Model and its Limitations

The cornerstone of risk management lies in accurate credit assessment, with credit risk being a primary concern for financial institutions. Since the 1950s, statistical credit scoring models (Hand & Henley, 1997) have been widely adopted[5].

These models typically employ techniques like linear regression and logistic regression to predict default probabilities based on borrowers' historical credit data and demographic information (Thomas, Edelman, & Crook, 2002)[6].

However, these traditional models exhibit significant limitations. They fail to effectively assess "credit white households" such as young individuals without credit history, rural populations, and new immigrants. Their heavy reliance on historical credit data has led to the "credit exclusion" phenomenon (Jagtiani & Lemieux, 2018)[7]. With low data update frequency and limited dimensions, traditional models predominantly rely on static existing information, making it difficult to capture dynamic changes in borrowers' financial conditions and repayment intentions. This fundamentally fails to address information asymmetry. In reality, default risks typically result from complex interactions of multiple factors, yet traditional models demonstrate relatively weak capabilities in capturing nonlinear relationships.

2.2 The Rise of Digital Footprints: as Supplementary Information for Credit Assessment

When individuals engage in online activities, the digital traces they intentionally or unintentionally leave behind are commonly defined as "digital footprints". These data sources are remarkably diverse, including but not limited to transaction records from e-commerce platforms, interaction data from social networks, mobile payment transaction histories, search engine query logs, and even geolocation data recorded by smartphone sensors (Oskarsdottir, Bravo, Sarraute, Vanthienen, & Baesens, 2019) [8]. Compared to traditional credit data, digital footprints demonstrate broader coverage, stronger real-time responsiveness, and more multidimensional characteristics, making them a highly promising "soft information" for credit assessment.

The value of digital footprints in credit assessment has begun to gain academic validation. A groundbreaking study by Berg et al. (2020) demonstrated that analyzing digital footprints of users on a German e-commerce platform—such as operating system type, email service provider, and typing speed—can build predictive models outperforming traditional credit scores, particularly effective in distinguishing "credit white users". Similarly, Varian (2014) noted that machine learning techniques can uncover complex patterns related to credit risk from massive search and transaction data [9]. These studies collectively conclude that digital footprints, which reveal user stability, integrity, and fulfillment capabilities at the behavioral level, can effectively supplement the limitations of traditional "hard information".

2.3 Risk Management Practices in Fintech Companies

Ant Financial, a flagship fintech company, stands as both a pioneer in implementing digital footprint-based risk management and a culmination of industry achievements. As China's first large-scale commercial application of digital footprint technology in personal credit assessment, its core product "Zhima Credit" evaluates users' creditworthiness through five dimensions: identity traits, fulfillment capacity, credit history, social connections, and behavioral preferences. Notably, data on behavioral preferences and social connections primarily originates from users 'digital footprints within the Alibaba ecosystem (Research Team of the Credit Reference Center of the People's Bank of China, 2017)[10].

When serving small and micro enterprises, Ant Group's MYbank has broken through traditional banking's risk control model that relies on financial statements and collateral. By analyzing merchants' transaction flows, store ratings, supply chain relationships, and even utility payment records on e-commerce platforms, MYbank conducts dynamic assessments of business operations and credit risks, achieving large-scale, low-cost, and efficient pure-credit lending services (Huang, Miao, & Wang, 2019)[11]. This risk control model based on digital footprints extends financial services to millions of small businesses that traditional banks struggle to reach, demonstrating significant inclusive value. More importantly, it enhances risk identification accuracy. However, the specific mechanisms by which digital footprints function in models and their comparative value increment to traditional financial features still require deeper empirical research for comprehensive analysis.

2.4 Research Design and Methodology

By constructing a data set containing the characteristics of traditional finance and digital footprint, this study uses simulation experiments to compare the performance of different models in the task of default prediction, so as to demonstrate the key value of digital footprint in risk management.

3 Research Design and Methodology

3.1 Build a Simulated Dataset

We designed and generated a simulated dataset containing 50, 000 user samples to replicate Ant Financial's data environment. Each sample represents a credit applicant with a set of features and a default label. In the real world, these features are distributed and correlated, and the dataset generation process aims to reflect these relationships [12].

3.2 Feature Definition and Generation Logic

The design of the feature set integrates two key dimensions: digital behavioral footprints and traditional financial hard data. Traditional features include annual income-to-debt ratio, credit history duration, and the number of credit accounts.

These indicators reflect users' financial status, debt repayment capacity, credit experience, and product usage breadth. Notably, the annual income is modeled as a log-normal distribution to better align with real-world patterns, while the debt-to-income ratio is designed to exhibit an inverse relationship with income levels. Behavioral data from payment e-commerce and social platforms serves as the primary focus for extracting digital footprint characteristics, encompassing three key dimensions: Transaction Activity – gauging user spending power through average transaction value and monthly transaction frequency; Online Behavior Patterns – capturing lifestyle preferences via shopping frequency and social media usage duration; User Engagement Patterns – assessing behavioral consistency and ecosystem loyalty based on stable utility payment scores and recent platform activity days. The default indicator, serving as the target variable in this study, is generated through logical synthesis of the aforementioned characteristics. A shorter credit history, higher debt-to-income ratio, and lower payment stability tend to increase default probability, while higher platform activity and income levels mitigate this risk. By applying the Sigmoid function and introducing random noise, we ultimately generated a sample set with a default rate of approximately 5%, which aligns with common practices in consumer credit [13].

3.3 Feature Processing and Filtering

All features underwent correlation screening and preprocessing prior to model training. Given that the target variable is binary, this study employed the point-bisector correlation coefficient to measure the association strength between continuous features and default status [14]. To ensure the selected features possess sufficient predictive power for subsequent modeling, a threshold of 0.05 was set for the absolute correlation coefficient, retaining only those exceeding this threshold for further modeling [15]. By eliminating weakly correlated variables, this step aims to emulate industry-standard procedures to enhance the model's generalization ability and training efficiency.

3.4 Model Building and Evaluation Strategies

This study developed two comparative models to evaluate the value of digital footprints in credit risk assessment. The benchmark model using traditional financial features employs logistic regression, demonstrating baseline performance of conventional risk control methods. The integrated model combining traditional and digital footprint features utilizes decision tree algorithm, which effectively captures variable interactions and nonlinear relationships in real-world risk scenarios, showing greater applicability. The tree structure provides excellent interpretability, and a pre-pruning strategy was implemented during training to control model complexity and prevent overfitting. Constraints and optimizations were applied to key hyperparameters including maximum depth and minimum split samples per node [16-17]. To comprehensively assess model performance, this study employs a series of metrics suitable for credit risk binary classification, including accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC) [18-19].

4 Empirical Results and Analysis

The value of digital footprints in credit risk assessment is systematically analyzed through simulated datasets and comparative model experiments, which falls within the scope of this chapter. The analysis reveals that digital footprints not only drive paradigm innovation in risk control models and provide unique risk signals, but also serve as a critical foundation for building core competitiveness in fintech platforms.

4.1 The Expansion of Risk Signals: From Financial Credit to Behavioral Credit

Through analyzing feature correlations to identify digital footprints, the study expanded risk assessment dimensions from "financial credit" to "behavioral credit". Research revealed a significant negative correlation between payment stability in daily expenses and default risk. This finding carries profound practical implications: When individuals consistently fulfill their payment obligations over extended periods, their behavioral patterns inherently demonstrate commitment to fulfillment, reflecting strong accountability and financial stability. Such persistent behavioral data that characterizes user creditworthiness at a fundamental level precisely constitutes the "soft information" missing in traditional credit reporting systems.

The number of active days on a platform is inversely related to default probability, driven by the "opportunity cost" mechanism. When deeply integrated users default, they lose their highly valuable digital identity, significantly reducing their incentive to breach ecosystem terms. Notably, not all digital behaviors carry equal predictive power—social app usage duration shows weaker correlation. This contrast highlights the critical need to accurately identify "strong signals" from massive digital footprints.

4.2 Innovation in Evaluation Paradigms: From Linear Evaluation to Dynamic Profiling

The clearly displayed model comparison results show that the introduction of digital footprint has driven a fundamental change in the credit risk assessment paradigm. It resolves the evaluation dilemma of "credit white users", a persistent challenge for traditional linear models. By shifting the evaluation basis from "past financial history" to "current behavioral patterns" through a decision tree model that integrates digital footprints, it creates technical possibilities for long-tail users overlooked by traditional financial services, thereby laying the foundation for financial inclusion.

To achieve more refined risk stratification, decision tree models can capture complex nonlinear relationships and interaction effects among features. These models may reveal a pattern: "When household utility payment stability falls below a specific threshold and the debt-to-income ratio is elevated, risks will surge dramatically. Risk assessment then transcends simple feature weighting, instead constructing multidimensional dynamic profiles based on actual behavioral patterns. This approach achieves a qualitative leap in both accuracy and robustness."

The "white-box" decision-making path provided by the decision tree model significantly enhances model interpretability. Risk control personnel can clearly trace the logical chain of each risk decision and understand the specific contributions of digital footprints. This not only facilitates deep integration of business processes and regulatory communication, but also transforms risk control from an "art of the black box" into an "interpretable science," making model auditing and optimization much more convenient.

5 Conclusion

This study discusses the effectiveness and logic of Ant Financial's use of digital footprint for risk control through simulation experiments, and obtains the following conclusions: First, digital footprints – as crucial "soft information" – contain user transaction data, payment habits, and platform engagement metrics that effectively reflect users' fulfillment intentions and life stability. These elements serve as vital supplements to traditional financial "hard information". Second, machine learning models integrated with digital footprints demonstrate significantly better risk prediction performance compared to linear models relying solely on conventional data. This model achieves more precise and interpretable risk assessments, not only capturing complex nonlinear risk patterns but also extending coverage to "credit white users". This study provides a clear reference for the digital transformation of traditional financial institutions at the practical level: it is necessary to actively integrate multi-source digital footprints, break the path dependence on traditional data and apply machine learning technology to restructure the risk control system.

This study has some limitations, such as the relatively simple simulation data model, and the lack of in-depth discussion on ethical issues such as data privacy and algorithmic fairness. Future research can be further expanded to try more complex models using real data, and strengthen the ethical regulation of digital footprint applications.

References

[1] Yao Zhao. "The Impact of Fintech on P2P Lending Risk and Credit Scoring." Statistics and Application [2024].

[2] Josef Scherer. "Compliance-) Risk Management System 4 . 0-The digital transformation of norms , guidelines and standards."[2019].

[3] Iyer R., Khwaja A. I., Luttmer E. F., & Shue K. Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2016). Screening peers in online credit markets: The role of signals. Journal of Financial Economics, 120(1), 43-66.

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

[5] Hand, W. Henley. "Statistical Classification Methods in Consumer Credit Scoring: a Review." Journal of the Royal Statistical Society: Series A (Statistics in Society) (1997).[1997-09-01].

[6] Thomas L. C., Edelman D. B., & Crook J. N. Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit scoring and its applications. SIAM.

[7] Jagtiani J. A., & Lemieux C. Jagtiani, J. A., & Lemieux, C. (2018). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. Financial Management, 48(4), 1009-1033.

[8] Óskarsdóttir M., Bravo C., Sarraute C., Vanthienen J., & Baesens B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26-38.

[9] Varian, H. R. Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28.

[10] Research Team of the Credit Reference Center, People's Bank of China. (2017). Comparative Study of Domestic and International Personal Credit Reporting Models. Credit Reference, 35(1),4-11.

[11] Huang Y., Miao K., & Wang Z.Huang, Y., Miao, K., & Wang, Z. (2019). Digital Finance and Corporate Innovation. NBER Working Paper No. 26571. National Bureau of Economic Research.

[12] K. Funatsu. "Knowledge-Oriented Applications in Data Mining." 2011.[2011-01-21].

[13] Laura Cleofas-Sánchez, V. Garcíaet al. "Financial distress prediction using the hybrid associative memory with translation." Appl. Soft Comput. (2016).[2016-07-01].

[14] Fatima Zahra Janane T. Ouaderhmanet al. "A filter feature selection for high-dimensional data." Journal of Algorithms & Computational Technology (2023).[2023-01-01].

[15] Ankur Jariwala, Aayushi Chaudhariet al. "Data Quality for AI Tool: Exploratory Data Analysis on IBM API." International Journal of Intelligent Systems and Applications (2022).[2022-02-08].

[16] Christophe Hurlin, C. Pérignonet al. "The Fairness of Credit Scoring Models." Microeconomics: Welfare Economics & Collective Decision-Making eJournal (2021).[2021-02-15].

[17] Sergio Rubio-Martín, María Teresa García-Ordáset al. "Enhancing ASD detection accuracy: a combined approach of machine learning and deep learning models with natural language processing." Health Information Science and Systems (2024).[2024-03-06].

[18] Chengyijing Wang, Haining Jianget al. "Customer Credit Rating by Machine Learning." BCP Business & Management (2023).[2023-01-13].

[19] Wanping Sun. "Research on the national income prediction based on Python." Applied and Computational Engineering (2023).[2023-10-23].