Consumption Behavior and Consumer Credit Risk: Heterogeneous Informational Value across Borrowers*

Jooha Nam[1] Jaejin Lee[2] Donghun Ryu[3] Younjoo Lee[4]

Abstract

We examine whether granular consumption behavior contains information about future default risk beyond conventional financial indicators and whether its value varies across borrowers. Using integrated credit and debit card transactions from Korea's MyData regime, we first estimate logit models and then nonlinear XGBoost models. Seven expenditure categories are significantly associated with default risk in the logit analysis. Consumption information provides greater incremental predictive value for lower-middle-income and younger borrowers, a pattern reinforced by XGBoost, SHAP-based importance measures, and bootstrap inference. These findings highlight meaningful heterogeneity in the credit-risk information contained in consumption behavior.

JEL codes: D12, D14, G21, G51.

---

* We are deeply grateful to Taehoon Kim, CEO of BankSalad, and the staff of BankSalad for providing the MyData and the empirical research environment used in this study. We thank Spencer Yongwook Kwon (Brown University), Ky-Hyang Yuhn (Florida Atlantic University), and Hojun Kang (Software Policy and Research Institute) for helpful comments and suggestions. We also thank Jun Hee Kwak, Sungbin Sohn, Rho-Sun Kwak, Taehyun Ahn, Yunjeong Kim, and Yoonho Jo for comments at the Sogang University seminar; Hosung Lim, Dongweon Lee, In Do Hwang, Taekyung Kim, Keunho Jang, Ji Eun Lee, Inro Lee, and Minyoung Lee at the Bank of Korea Economic Research Institute seminar; Yongki Hong at the annual meeting of the Korean Money and Finance Association; and Tae-Rog Oh, Jinsoo Bae, Bon-Sung Gu, Mansoo Jee, Kyoobok Lee, and Sungwook Park at the Korea Institute of Finance seminar. The findings and views expressed in this paper are those of the authors and do not represent the official position of BankSalad, the data provider. All remaining errors are our own.

[1] Professor Emeritus, School of Economics, Sogang University. E-mail: namjh0506@gmail.com.

[2] Assistant Professor of Finance, Kennesaw State University. E-mail: jlee589@kennesaw.edu. Corresponding author.

[3] Ph.D. candidate, Department of Economics, Sogang University. E-mail: ryudh@ii.re.kr.

[4] Head of Corporate Strategy, BankSalad. E-mail: younjoo.lee@banksalad.com.

Assessing individual credit risk is central to the soundness of commercial banks and the efficient allocation of credit, yet lenders largely rely on borrower-level indicators such as income, savings, assets, and debt. These variables summarize financial capacity but reveal relatively little about how borrowers allocate resources in everyday life. Granular consumption behavior can capture preferences, constraints, and behavioral tendencies that are only imperfectly reflected in conventional financial information. An important question is therefore not only whether transaction-level consumption data predict future default, but also for which borrowers such information is most valuable.

This paper addresses both questions. We first examine whether the composition of granular consumption expenditure contains information about future consumer default risk beyond standard financial and demographic controls, and which expenditure categories carry that information. We then make borrower heterogeneity a central part of the analysis by asking whether the incremental predictive value of consumption differs across income quintiles and age groups. This second question is economically important because the marginal value of alternative information should be greatest when conventional financial variables provide relatively limited discrimination.

The heterogeneity perspective changes how consumption information should be evaluated in credit assessment. A small average improvement in predictive performance can mask economically meaningful gains for particular borrower groups. Income and age are especially relevant dimensions because they are closely related to the quantity and informativeness of conventional financial information and to differences in consumption patterns. We therefore estimate the models separately across five income quintiles and five age groups and compare the incremental contribution of the same disaggregated consumption variables within each subgroup.

Empirical evidence on these questions has been limited by data access. Prior studies often rely on surveys, field experiments, or transaction records from a single lender or payment provider, which provide only a partial view of individuals' spending. Korea's MyData regime, introduced in 2022, provides an institutional framework for integrating transaction-level financial information across institutions with individual consent. We use actual credit and debit card transactions collected through BankSalad. Our final sample contains 202,247 borrowers, with a 3.1% default rate over the one-year prediction window from February 2025 to January 2026. The data allow us to organize card spending into 15 disaggregated categories and relate consumption observed during the preceding year to subsequent default.

Our empirical strategy combines two complementary approaches. We first estimate logit models, which provide an interpretable benchmark for identifying the direction, magnitude, and statistical significance of the association between category-level consumption behavior and future default. We also use the logit framework to examine whether the incremental discriminatory power of consumption varies across income quintiles and age groups. The subgroup estimates indicate that consumption information is more informative among lower-middle-income borrowers and tends to contribute more among younger borrowers. We then use XGBoost as a complementary nonlinear specification that relaxes the functional-form restrictions of the logit model and allows for nonlinearities and interactions. SHAP values and permutation-based AUROC drops are used to assess whether the importance of the seven consumption variables identified by the logit analysis is preserved across borrower groups, while paired bootstrap inference evaluates the precision of the incremental predictive gains.

Three results stand out. First, consumption behavior contains information about future default beyond standard financial and individual characteristics. In the logit model, seven expenditure categories are statistically significant: telecommunications, convenience-store, and café-and-snack spending are positively associated with default risk, whereas education, sports and fitness, apparel, and medical and health spending are negatively associated. Second, the informational value of consumption is heterogeneous across income groups. In the logit subsample analysis, adding consumption variables produces the largest improvements in discrimination in the second and third income quintiles (incremental test-sample AUROC of +0.0318 and +0.0206, against +0.0031 for the full sample), and bootstrap inference distinguishes these gains from zero. XGBoost yields the same broad pattern, with particularly large incremental gains in these lower-middle-income groups (+0.0443 and +0.0370, against +0.0179 for the full sample). SHAP importance and permutation-based AUROC-drop measures further show that the seven significant consumption variables carry substantial predictive information in these groups.

Third, the age-group analysis also indicates that the informational contribution of consumption is concentrated among younger borrowers. In the logit model, the incremental AUROC is largest for borrowers aged 29 or below (+0.0033) and declines across the older age groups (to +0.0009 for borrowers in their fifties), although the age-specific logit increments are estimated less precisely. The complementary XGBoost analysis strengthens this pattern: the youngest group shows the largest predictive gain (+0.0465, compared with +0.0183 for borrowers in their thirties and +0.0050 for those aged 60 or above). SHAP importance and permutation-based AUROC-drop measures point in the same direction. Taken together, the logit and XGBoost evidence suggests that consumption information is especially useful for assessing credit risk among lower-middle-income and younger borrowers, while the nonlinear model reveals additional predictive structure not fully captured by the logit specification.

These findings contribute to the literature in three ways. First, we use comprehensive MyData-based card-spending records that consolidate transactions across financial institutions, allowing consumption behavior to be observed at a level of coverage unavailable in most lender-specific datasets. Second, we identify seven disaggregated expenditure categories that contain information about future default beyond conventional financial controls. Third, and most importantly, both the interpretable logit analysis and the complementary XGBoost analysis point to substantial heterogeneity in the informational value of consumption across borrowers. The evidence is strongest for lower-middle-income borrowers and younger borrowers, and is reinforced by subgroup prediction, SHAP importance, permutation-based AUROC drops, and bootstrap inference. The central implication is economic rather than technical: granular transaction information is not equally informative for all borrowers, and its value is concentrated in groups for which conventional financial information may provide less complete discrimination of credit risk.

Our paper relates to three strands of the literature. The first links credit risk to the composition of consumption and debt. Zhu (2011) combines Delaware personal bankruptcy filings with the Survey of Consumer Finances and finds that higher ratios of mortgage, auto loan, and credit card debt to income, as well as medical expense burdens, are associated with a higher probability of bankruptcy. Kim, Cho, and Ryu (2018) estimate delinquency-state transition probabilities on Korean credit card loan accounts and find that higher DTI ratios and larger remaining balances raise the probability of entering delinquency. The most closely related paper is Bakhtiari et al. (2026), who analyze account data from a single large U.S. credit card issuer with a hidden Markov model and document that category-level spending patterns predict credit card delinquency: spending at discount and second-hand stores and on entertainment and nightlife is positively related to delinquency risk, whereas spending on necessities such as groceries and gasoline is negatively related. To the best of our knowledge, prior research has not examined the relationship between granular consumption behavior across a comparably broad range of expenditure categories and subsequent credit risk using integrated transaction-level records such as those available through Korea's MyData regime. More importantly, we move beyond average predictive performance to examine whether the incremental informational value of consumption varies systematically across borrowers. We return below to how our data and categories differ from theirs.

The second strand argues that spending behavior reveals intrinsic borrower traits that lenders cannot observe directly. Karlan and Zinman (2009) separate hidden information from hidden action in a South African consumer credit field experiment and attribute approximately 13% to 21% of default to moral hazard. Vissing-Jorgensen (2021) relates indicators for 441 detailed consumption categories to repayment health in U.S. and Mexican data, finds that entertainment-related items such as lotteries and pet spending consistently predict deteriorating repayment health, and interprets the association as originating in impatience.[1] A related behavioral literature ties over-indebtedness to psychological and cognitive traits: Laibson (1997) formalizes present bias through hyperbolic discounting; Gathergood (2012) finds that lack of self-control explains over-indebtedness better than poor financial literacy; Meier and Sprenger (2010) show that more present-biased individuals hold more credit card debt; Kuchler and Pagel (2021) show that present-biased individuals repeatedly fail to follow through on credit card paydown plans; Gong et al. (2024) show that impulsiveness proxied by gaming expenditure

1 There is also an active literature on alternative information sources distinct from traditional credit-assessment inputs such as credit history and income. Berg et al. (2020) use transaction data from a German e-commerce firm and show that the digital footprint users leave during website access and registration (device type, operating system, e-mail provider, and the like) carries informational value for default prediction comparable to credit bureau scores, and that the two act as complements.

predicts higher post-issuance default; and D'Acunto et al. (2023) show that individuals with higher cognitive ability make smaller inflation forecast errors and formulate more accurate consumption plans.

These prior studies share a common limitation in how consumption is measured. Gathergood (2012) and Meier and Sprenger (2010) document that self-control and present bias shape debt behavior, but their evidence rests on surveys or limited field experiments rather than on actual payment records. Gong et al. (2024) use actual transaction data but take a single item, gaming expenditure, as a proxy for impulsiveness.

The third strand studies alternative information and nonlinear methods in credit assessment. Khandani, Kim, and Lo (2010) combine customer transaction and credit-bureau data from a large commercial bank and show that machine-learning models materially improve forecasts of consumer delinquency and default. Jagtiani and Lemieux (2019) document that alternative information used by a fintech lender adds predictive content beyond traditional credit scores, while Berg et al. (2020) show that digital footprints provide default-prediction information that complements conventional credit-bureau measures. More recently, Di Maggio and Ratnadiwakara (2026) show that alternative data can be particularly valuable for 'invisible primes,' borrowers whose traditional credit records understate their credit quality. These studies establish the value of nontraditional information for credit assessment. Our contribution is to identify the economic content of granular consumption signals and to show that their incremental informational value varies systematically across borrower income and age.

1 Data and Institutional Background

1.1 The MyData regime

Our analysis exploits detailed individual-level financial transaction data made available through Korea's MyData regime, introduced in 2022. The regime enables individuals to authorize the collection and integration of their financial information across institutions, allowing their financial activities to be viewed across multiple institutions. Much of the existing literature relies on conventional household surveys, credit bureau data, or field experiments, which provide valuable but often limited information on individuals' financial behavior and outcomes. In contrast, our data contain detailed information on individuals' financial transactions and account balances over time.

Korea's MyData regime builds on the country's broader open-banking infrastructure but goes beyond the simple aggregation of account information. Under the regime, individuals can authorize licensed MyData providers to collect and consolidate their financial information from multiple financial institutions in a standardized format and access it through a single service. The information can include bank accounts, cards, loans, investments, and other financial products, thereby reducing the fragmentation of financial information across institutions. Importantly for empirical research, the regulatory framework also permits licensed MyData providers to use or provide appropriately processed data for permitted purposes, including research, subject to applicable legal and privacy requirements.

Korea's extensive use of bank accounts, payment services, and mobile financial platforms provides a particularly rich empirical environment for studying individual consumption behavior.

The data allow us to construct detailed measures of consumption from financial transactions and to examine whether these behavioral patterns contain information about subsequent credit risk. [2]

Our sample covers individuals for whom we observe monthly financial transactions and account-level information over the sample period. We aggregate transaction-level records into individual-month measures of consumption and combine these measures with information on subsequent credit outcomes. This structure enables us to examine not only the level of consumption but also its composition and dynamics over time, providing a richer characterization of individual consumption behavior than is available from aggregate expenditure measures.

1.2 Sample construction and variable definitions

Our sample starts from 1,003,293 BankSalad users with transaction records during February 2024–January 2026 who were not in default at the beginning of the default observation window. This restriction allows us to examine new default events among borrowers who were non-default at the prediction point. We construct the analysis sample in four steps. Appendix Table A1 reports the number of users and the default rate at each step.

In the first step, we retain borrowers with observed debt. Our outcome is a default event on outstanding debt, so the outcome is not defined for a user who holds no loan. We therefore exclude 640,261 users whose loan balance is zero, negative, or unobserved. This step leaves 363,032 users, of whom 10,390 default, a default rate of 2.86%.

[2] The data used in this study were provided by BankSalad, a licensed MyData provider, to the research team at Sogang University for research purposes under the applicable legal and regulatory framework governing the use and provision of financial data. BankSalad has been authorized to conduct data-related ancillary business, including the provision and intermediation of appropriately processed data, under Korea's Credit Information Use and Protection Act and related regulations. BankSalad classifies card payments at the merchant level into 29 disaggregated categories comprising 119 sub-items.

In the second step, we retain borrowers with observed income. Every consumption variable is a ratio of category spending to income, and log income and the financial-assets-to-income ratio enter the model as financial controls; none of these is defined when income is recorded as zero. In addition, under Korea's MyData framework, records from a financial institution are transmitted only for accounts that the user has explicitly consented to link, and income is observed only when the account through which salary or other income is received is among them. A recorded income of zero may therefore reflect incomplete linkage of the account through which salary or other income is received rather than genuinely zero income. These users hold observed loans and record card spending during the observation window, suggesting that at least some recorded zeros may reflect incomplete linkage rather than an actual absence of income.

Of the 169,419 users whose total income is recorded as zero, 139,733 have already been removed in the first step, so this step excludes a further 29,686 users. It leaves 333,346 users, of whom 8,727 default, a default rate of 2.62%.

In the third step, we trim extreme values. We drop 19 users with missing individual characteristics such as age, and we then trim both tails of each continuous variable by approximately 3%. Because the trimming is applied variable by variable, a user is removed if any single variable takes an extreme value, which excludes a further 66,666 users. This step leaves 266,661 users, of whom 6,258 default, a default rate of 2.35%.

In the fourth step, we apply stratified sampling. After the first three steps the default share is 2.35%, which leaves the estimation sample unbalanced between defaulted and non-defaulted borrowers. Classification models trained on such a sample tend to fit the minority class poorly, and the default share is also low relative to the segment of the credit market in which alternative information is most likely to be used. We therefore rebalance the sample in one direction only.

We retain all 6,258 defaulted borrowers and draw a random sample of the non-defaulted borrowers, reducing that group from 260,403 to 195,989. Because no defaulted borrower is dropped, the procedure discards no information about default events themselves; it removes only non-defaulted borrowers.

The sample of non-defaulted borrowers is stratified by gender and age group. These two variables enter the model as individual characteristics and are related to both consumption patterns and credit risk, so unstratified sampling could shift the demographic composition of the non-default group and confound its comparison with the default group. Stratification holds that composition fixed. The resulting analysis sample consists of 202,247 individuals, 195,989 non-defaulted and 6,258 defaulted, with a default rate of 3.09%. AUROC[3], the Gini coefficient, and the K-S statistic are rank-based measures and do not depend on the share of defaulted borrowers in the sample on which they are computed, so the rebalancing does not mechanically affect the main performance measures. The PR-AUC does depend on that share and is accordingly interpreted through its improvement over the baseline model rather than against a fixed threshold.

We split the final sample 80:20 into a training sample of 161,797 users and a test sample of 40,450 users, maintaining a similar default rate in each.

We set January 2025 as the credit risk prediction point. The preceding 12 months, February 2024–January 2025, constitute the financial observation window, while the dependent variable indicates whether a new default occurs during the subsequent 12 months, February 2025–January 2026. Using equal-length observation and prediction windows allows us to relate recently

[3] We retain all defaulted borrowers and subsample only the non-default group to preserve information on relatively rare default events while maintaining a sufficiently large comparison group. This procedure results in a default share of 3.09% in the final analysis sample. The sampling procedure and the characteristics of excluded borrowers are reported in Appendix Table A1.

observed consumption behavior to subsequent credit risk while reducing the influence of seasonal variation.

[Figure 1 about here.]

Following the criteria used by Korea Credit Bureau and other credit information institutions, we define credit risk as the occurrence of a default event with a lasting effect on credit status, including debt nonperformance, subrogation, bankruptcy, debt restructuring, and relevant public records. We exclude short-term delinquencies that are more likely to reflect temporary liquidity constraints and, when multiple default events occur, use the first event in the borrower's lifetime.

Our explanatory variables measure consumption behavior using spending-to-income ratios. For each category, cumulative spending during the 12-month observation window is divided by cumulative income over the same period. This normalization reduces differences attributable to absolute income levels and allows us to characterize differences in consumption patterns across individuals. Of the 29 disaggregated expenditure categories, we exclude nine, which fall into three types. First, categories whose purchased items are so heterogeneous that consumption behavior cannot be characterized: online shopping, general shopping, and department stores, outlets, and malls. Second, financial and administrative outlays that reflect institutional obligations rather than consumption behavior: insurance and taxes, and online payments. Third, categories for which the economic interpretation of the spending motive is ambiguous or whose contents mix spending of different natures: automotive, overseas payments, daily living, and other. The twenty categories that remain are reduced to nineteen by combining the bar and nightlife categories, which are similar in nature. These nineteen disaggregated categories form the basis of the subsequent empirical analysis.

1.3 Summary statistics

Table 1 presents summary statistics for the main variables for the non-default group (195,989 individuals) and the default group (6,258 individuals, approximately 3.1% of the sample). Financial controls include log income, the ratio of financial assets to income, and the ratio of debt to financial assets. For ease of interpretation, income is reported in units of 10,000 KRW. Consumption expenditure variables enter as raw spending-to-income ratios without additional standardization.

[Table 1 about here.]

The two groups differ systematically in their consumption patterns. Defaulted borrowers exhibit higher spending-to-income ratios in several convenience-oriented categories, including convenience stores, telecommunications, food delivery, and taxis. Medical and health spending is also somewhat higher among defaulted borrowers.

The financial controls show even larger differences. The debt-to-financial-assets ratio is substantially higher among defaulted borrowers (1,134.4 versus 337.2), while the financial-assets-to-income ratio is considerably lower (0.110 versus 0.304). Income is also lower among defaulted borrowers (3,542 versus 5,446, in 10,000 KRW). These differences indicate that defaulted borrowers have weaker financial positions overall. Because the univariate comparisons do not account for differences in borrowers' financial and demographic characteristics, we examine whether the consumption patterns remain associated with credit risk after controlling for these factors in the multivariate analysis.

2 Consumption Behavior and Default Risk: Logit Evidence

2.1 Univariate evidence

We begin with independent two-sample t-tests of whether the disaggregated consumption-behavior variables, each measured as the ratio of category spending to income, discriminate between the non-default and default groups within one year of the observation date (Table 2). Among the 19 disaggregated expenditure categories, the mean difference is statistically significant for 15; the exceptions are travel, public transit, childcare, and housing. Across the significant categories, the default group exhibits higher spending-to-income ratios across the board rather than excessive spending concentrated in particular categories. The largest differences belong to telecommunications, convenience stores, food delivery, and taxis: the telecommunications spending-to-income ratio averages 0.263 in the default group versus 0.090 in the non-default group, roughly 2.9 times higher, and convenience-store spending is about 2.0 times higher (0.083 versus 0.042). The gaps are thus most pronounced in categories characterized by the pursuit of immediate convenience and by inelastic demand.

[Table 2 about here.]

Medical and health spending also differs significantly, with a higher mean in the default group (0.154) than in the non-default group (0.128). These univariate comparisons call for caution, however: we cannot rule out the possibility that they reflect a structural feature of ratio variables, whose values tend to be high among low-income groups through the denominator. The financial control variables display clear between-group differences consistent with theoretical predictions. The financial assets-to-income ratio shows the largest difference of all variables and is much lower in the default group (0.110 versus 0.304); the debt-to-financial assets ratio is about 3.4 times higher in the default group (1,134.4 versus 337.2); and income, in units of KRW

10,000, is higher in the non-default group (5,446.0 versus 3,541.9). Because correlations among the variables, and in particular the tendency of ratio variables to be high among low-income borrowers, limit what univariate comparisons can identify, the analyses that follow estimate logit models by maximum likelihood that include the financial controls and individual characteristics simultaneously, thereby isolating the independent contribution of each consumption behavior.

2.2 Empirical specification

We estimate the following specification:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \sum_{j\in J_m} \beta_j\ X_{ij} + \sum_{f=1}^{3} \gamma_f\ F_{if} + \delta\,\text{Gender}_i + \sum_{d=1}^{4} \lambda_d\ \text{Age}_{id}, \qquad (1)$$

where P_i denotes the probability of default for borrower i; X_{ij} is borrower i's spending-to-income ratio for expenditure category j, where the set of categories is

$$J_m = \begin{cases} J_C = \{j_{c_1}, j_{c_2}, \dots, j_{c_{19}}\}, & \text{if } m = C \text{ (the disaggregated model),} \\ J_K = \{j_{k_1}, j_{k_2}, \dots, j_{k_8}\}, & \text{if } m = K \text{ (the capital-type model);} \end{cases}$$

F_{if} is the f-th financial control variable for borrower i, with $f = 1, 2, 3$ (log income, financial assets to income, and debt to financial assets); Gender_i is an indicator variable that equals one if borrower i is female and zero otherwise; and Age_{id} are age-group dummies with $d = 1, \dots, 4$ (borrowers in their thirties, forties, fifties, and sixties or above), where borrowers in their twenties and younger form the reference group. That is, X_{ij} covers the disaggregated spending-to-income ratio variables, measured as ratios to annual income rather than raw expenditure levels to control for heterogeneity in income scale. The disaggregated specification initially includes the 15 of the 19 expenditure categories for which the mean-difference tests indicate a significant difference between the default and non-default groups (Table 2), allowing us to identify the independent association of each expenditure category with default risk. In the subsequent risk-factor specification, we retain the seven consumption variables that remain statistically significant after the full set of financial and individual controls is included. The three financial controls capture the basic pillars of individual credit assessment: log income measures the borrower's absolute income level (log-transformed to mitigate the right skewness of the income distribution), financial assets to income measures accumulated asset capacity, and debt to financial assets measures the debt burden, with the latter two constructed as ratios rather than absolute amounts to control for heterogeneity in income and asset scale across borrowers.

Multicollinearity is not a concern: variance inflation factors (VIFs) are below 2 for every disaggregated consumption variable (1.03 to 1.74), well under the conventional threshold of 5, and pairwise correlations are likewise modest (see Appendix Table A2). The highest correlations among the disaggregated variables, around 0.4, form a cluster among the immediate-convenience channels (Taxis, Food delivery, Convenience stores, and Restaurants), and log income is negatively correlated with most expenditure ratios, at about 0.4 with Public transit, Food delivery, Convenience stores, and Restaurants. These patterns are consistent with the structural feature of spending-to-income ratios whereby expenditure shares tend to rise as income falls.

We first examine models that exclude the financial controls, and we then estimate the models that include them sequentially in six steps, varying the set of controls. Column 1 is the income baseline model, which includes only log income and the individual characteristics. Column 2 adds the complete set of consumption variables to column 1. Column 3 is the financial baseline model, which adds financial assets to income and debt to financial assets so that it contains the full set of financial controls together with the individual characteristics, and serves as a benchmark for the predictive power of the existing credit assessment framework. Column 4 is the full model, which adds the complete set of consumption variables to column 3. Column 5 is the risk-factor model, which applies stepwise selection and retains only the variables statistically significantly related to default risk, and column 6 reports the average marginal effects (AMEs) computed from the column 5 estimates. This sequential design isolates the incremental predictive power that the consumption variables contribute independently of the existing financial information, measured by the improvement in the test-sample area under the receiver operating characteristic curve (AUROC)[4] (ΔTest AUROC).

Setting two baselines separates the incremental predictive power of consumption from the definition of the benchmark, because the increment the consumption variables add depends on the scope of information included in the baseline: contrasting columns 1 and 2 measures the increment over income alone, and contrasting columns 3 and 4 measures it over the full financial information set. As a benchmark, an AUROC above 0.60 in information-scarce environments and above 0.70 in information-rich environments is known to indicate sufficient discriminatory power (Iyer et al. 2015).

For an intuitive interpretation of the estimated coefficients, we also present odds ratios (ORs), $\mathrm{OR}_{j_m} = e^{\hat{\beta}_{j_m}}$. Because the consumption variables are measured as raw spending-to-income ratios with values between zero and one, $\exp(\beta)$ itself corresponds to the change in the odds when spending increases by the full amount of income, an excessive change rarely observed in actual consumption distributions. We therefore anchor the interpretation on a 1-percentage-point increase in the expenditure share of income: if the estimated coefficient on the

[4] The AUROC is a standard measure of the overall discriminatory power of a classification model. When the model analyzes each borrower's consumption pattern and ranks borrowers from most to least risky, the AUROC is the probability that the model correctly places an actual defaulted borrower at a riskier (higher) rank than a non-default borrower. A value closer to one means that the model separates risky from safe borrowers perfectly, whereas 0.5 corresponds to a coin flip, that is, a purely random ranking with no informational content. For example, an AUROC of 0.70 means that if defaulted and non-default borrowers are randomly paired one to one, the model assigns the higher risk score to the defaulted borrower in about 70 of 100 pairs.

telecommunications expenditure ratio is $\beta = 0.155$, a 1-percentage-point increase in the telecommunications share of income yields an odds ratio of $\exp(0.01 \times 0.155) \approx 1.0016$, so the default odds rise by about 0.16%; a negative coefficient implies a symmetric decrease, and an odds ratio of one means that the expenditure ratio is unrelated to default risk. To measure the absolute change in the default probability itself, we compute the AME,[5] which expresses, in percentage points, the average change in a borrower's default probability when the corresponding expenditure share of income increases by 1 percentage point.

2.3 Multivariate evidence: Disaggregated-category results

We first present estimates based only on the individual characteristics and consumption variables, excluding the financial controls, to establish how well consumption behavior alone predicts default risk in the absence of financial information (Table 3). Column 1 includes only the full set of consumption variables, column 2 adds the individual characteristics, and column 3 retains only the statistically significant consumption variables from column 2, and column 4 reports the average marginal effects of the column 3 specification; the discussion focuses on columns 3 and 4.

[Table 3 about here.]

The test AUROC is 0.6508 in column 1, 0.6591 in column 2, and 0.6618 in column 3, confirming that individual characteristics and consumption behavior alone deliver predictive power of about 0.66, and the variables that differ significantly in the mean-difference tests (Table 2) are generally confirmed in the same direction. Based on column 3, the variables significantly

[5] In the logit model the default probability p is obtained by passing the linear combination of covariates $X\beta$ through the logistic function, $p = 1/(1 + e^{-X\beta})$. The marginal effect of a variable x_j is therefore the derivative $\partial p / \partial x_j = p(1-p)\beta_j$, which depends not only on the regression coefficient β_j but also on the level of the default probability p at that point. We compute $p(1-p)\beta_j$ at each borrower's own predicted default probability for every borrower in the sample and then average, which yields the AME. It expresses the average change in the default probability, in percentage points, when the expenditure ratio increases by one unit.

positively related to default risk are convenience stores, telecommunications, taxis, food delivery, culture and leisure, and cafés and snacks (coefficients of 0.680, 0.246, 0.241, 0.186, 0.170, and 0.076), while sports and fitness and medical and health are significantly negative (−0.296 and −0.092). That the immediate, convenience-oriented categories are consistently positive even without financial information shows that consumption behavior itself contains an independent predictive signal for default risk. In terms of the column 4 AMEs, a 1-percentage-point increase in the corresponding expenditure share raises the default probability on average by 0.0202 percentage points for convenience stores, 0.0073 percentage points for telecommunications, and 0.0072 percentage points for taxis, and lowers it by 0.0088 percentage points for sports and fitness. Among the individual characteristics, the dummies for ages 50–59 and 60 or above are significantly positive (coefficients of 0.274 and 0.897) at the 1% level relative to borrowers in their twenties, whereas the dummies for ages 30–39 and 40–49 and the gender indicator are insignificant.

We next add the financial controls (Table 4). The income baseline model in column 1, which includes only log income and the individual characteristics, attains a test AUROC of 0.6523; adding the full set of consumption variables in column 2 raises it to 0.6664, an improvement of 0.0141. The financial baseline model in column 3, which includes the full set of financial controls, attains a test AUROC of 0.7551; adding the consumption variables in column 4 raises it to 0.7582, an improvement of 0.0031, and the risk-factor model in column 5 maintains a similar level at 0.7577. The pseudo-R^2 rises from .0759 in column 3 to .0791 in column 4.

Although the same consumption variables are added, the improvements over the two baselines differ, at 0.0141 and 0.0031, a difference that stems from the scope of information included in the baseline. Because financial assets to income and debt to financial assets are variables with income in the denominator, including them in the baseline means that part of the information contained in the consumption variables is already reflected at the baseline stage. That adding the consumption variables raises predictive power in a consistent direction even with financial information controlled for this broadly shows that consumption information contains an independent signal that financial and individual characteristics alone do not capture.

The coefficient interpretation that follows is based on the risk-factor model in column 5, which includes the full set of financial controls: this takes the information set used by the existing credit assessment framework as the baseline and measures the additional contribution of consumption information conservatively. The gap between the training-sample AUROC (0.7626) and the test-sample AUROC (0.7582) in column 4 is only 0.0044, confirming stable generalization rather than overfitting.

[Table 4 about here.]

Based on the risk-factor model in column 5, 7 of the 15 entered consumption variables are statistically significantly related to default risk: convenience stores, telecommunications, and cafés and snacks positively (coefficients of 0.187, 0.154, and 0.057), and education, sports and fitness, apparel, and medical and health negatively (−0.307, −0.293, −0.194, and −0.160). When the convenience-store share of income increases by 1 percentage point, the default odds rise by about 0.19% and the default probability increases by 0.0054 percentage points on average; the corresponding figures are about 0.15% and 0.0044 percentage points for telecommunications and about 0.06% and 0.0017 percentage points for cafés and snacks. The relatively large marginal effects for convenience-store and telecommunications spending suggest that even routine, individually modest expenditures may contain meaningful information about credit risk. Their predictive content may arise not from the economic significance of individual purchases, but from the broader spending patterns they reveal about borrowers, patterns that conventional balance-sheet and income measures may not fully capture.

In contrast, a 1-percentage-point increase in the education expenditure share lowers the default odds by about 0.31% and the default probability by 0.0089 percentage points; sports and fitness (0.29%; 0.0085 percentage points), apparel (0.19%; 0.0056), and medical and health (0.16%; 0.0046) move in the same direction. The negative direction of education spending suggests a disposition toward continued self-development and relative stability in labor productivity and repayment capacity, in line with the view of education and training expenditures as human capital investments that raise future productivity and wages (Becker 1962); that of medical and health spending is consistent with health capital theory, whereby sustained health maintenance prevents interruptions to labor supply and preemptively mitigates liquidity pressure from unexpected health shocks (Grossman 1972). That these negative associations survive after controlling for income and financial assets suggests that they do not simply stem from differences in income levels.

The expenditure ratios for culture and leisure, beauty and personal care, taxis, pets, food delivery, restaurants, bars and nightlife, and gifts and family occasions are statistically insignificant once the financial variables are controlled for. Taxi, food-delivery, and culture-and-leisure spending were significantly positive in Table 3 but lose significance once income, financial assets, and the other financial variables are included: much of their explanatory power is absorbed by their correlation with income levels, so these expenditures appear closer to proxy indicators of income level than to independent risk signals in their own right.

The financial controls remain significant at the 1% level in all specifications. A 1% increase in income lowers the default odds by about 0.49% and the default probability by 0.0141 percentage points on average; a 1-percentage-point increase in the financial-assets-to-income ratio lowers the default odds by about 1.6% and the default probability by 0.0458 percentage points, the largest risk-mitigating effect among the financial controls. Because the debt-to-financial-assets ratio is measured on a markedly different scale, we interpret only its sign, which is significantly positive, consistent with heavier debt burdens raising default risk. The financial-structure variables thus retain stronger predictive power than the individual consumption variables, but the consumption variables add an improvement the financial variables alone do not explain. Among the individual characteristics, borrowers in their thirties, fifties, and sixties and above have default odds about 18%, 27%, and 82% higher than the reference group in their twenties, all significant at the 1% level; the dummy for ages 40–49 is insignificant. The dummy for ages 30–39, negative and insignificant without the financial variables (Table 3), turns positive and significant once income and assets are controlled for, implying that default risk within an age group can look different once income levels are held fixed. Female borrowers' default odds are about 5% lower than male borrowers', significant only at the 10% level.

Because the log-income control imposes a log-linear relation between income and default risk, we also re-estimate the model replacing log income with income quintile dummies; full estimates are reported in Table 5. With the lowest quintile as the reference group, the default odds decline monotonically but concavely in income: the coefficient falls steeply from 0.169 for Q20–Q40 to 0.864 for Q40–Q60 and 1.505 for Q60–Q80, whereas the additional decline shrinks markedly at Q80 and above, at 1.664. Income gains from low to middle income thus substantially lower default risk, whereas the risk-mitigating effect diminishes beyond a certain level. The consumption-variable estimates are broadly consistent with Table 4: the positive variables (telecommunications, cafés and snacks) and the negative variables (education, sports and fitness, apparel, medical and health) retain their signs and significance. Some associations sharpen, however: the convenience-store coefficient increases from 0.187 (column 5 of Table 4) to 0.442 (column 3 of Table 5), and taxi spending, insignificant in the log-income model, becomes significantly positive in the income-quintile model (coefficient 0.300), implying that the associations of some expenditure categories can emerge more clearly when income is controlled for more flexibly with quintile dummies.

[Table 5 about here.]

3 Machine Learning Evidence: XGBoost and SHAP

3.1 Methodology

To complement the logit analysis, we next examine the relation between the spending-to-income variables and borrower credit risk using XGBoost, a gradient boosting decision tree algorithm (Chen and Guestrin 2016). The two approaches play distinct roles. The logit model permits direct interpretation of the sign, magnitude, and statistical significance of the coefficients on the category-level spending-to-income ratios and is therefore well suited to testing economic hypotheses about the direction in which consumption structure affects default risk. However, the pairwise correlations among the disaggregated expenditure category variables are generally low (at most 0.4), and the relation between expenditure shares and default risk may be nonlinear, with default risk rising sharply once a threshold level is exceeded. Lundberg et al. (2020) show that, when such nonlinearity is present, linear models can distort not only prediction errors but also the interpretation of variable effects: the stronger the nonlinearity, the more the coefficient estimates may be biased.

We therefore use XGBoost for two complementary purposes. The first is to assess whether allowing for nonlinearities and interactions improves out-of-sample predictive power beyond the logit specification. The second is to examine whether the seven spending-to-income variables that remain significant in the fully controlled logit model continue to display economically consistent direction and importance when functional-form restrictions are relaxed. If the signs of the logit coefficients agree with the direction implied by the SHAP-based diagnostics and the variables retain meaningful importance in XGBoost, this provides evidence that their informational content is not specific to a particular model structure. The interpretive difficulty created by XGBoost's black-box structure is, in turn, mitigated by decomposing each expenditure variable's contribution to the model's predictions using the SHAP methodology described below. This approach follows research practice in the credit risk literature, including Bussmann et al. (2021) and Chen, Calabrese, and Martin-Barragan (2024).

XGBoost is an additive model that sequentially grows K decision trees (f_k) and sums their outputs, so that the default risk score $\hat{y}_i$ of borrower i is defined as

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{K} f_k\,(x_i), \qquad f_k \in F. \qquad (2)$$

The model is estimated sequentially to minimize an objective function that jointly controls prediction error and structural complexity. The objective function at the t-th iteration, $\mathcal{L}^{(t)}$, is defined as

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{k=1}^{t} \Omega\,(f_k), \qquad \Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2\,, \qquad (3)$$

where l is the logistic loss appropriate for a binary dependent variable and $\Omega(f_t)$ is a regularization penalty imposed on the number of leaves T and the leaf-weight vector w_j. The penalty parameters γ (the minimum loss reduction required for a split) and λ (the L2 regularization coefficient) curb the overfitting that can arise from the set of disaggregated consumption expenditure variables, thereby improving out-of-sample predictive performance.

We optimize the hyperparameters using Optuna, a Bayesian optimization framework, and apply stratified five-fold cross-validation to address the imbalance between defaulted and non-defaulted borrowers in the sample. The final hyperparameters applied to the disaggregated XGBoost model are reported in Appendix Table A3.

Because the XGBoost model combines variables through complex nonlinear interactions, the influence of an individual expenditure variable cannot be interpreted directly in the way a logit regression coefficient can. To address this, we apply Shapley Additive exPlanations (SHAP; Lundberg and Lee 2017), which builds on the theoretical foundation of Shapley (1953) in cooperative game theory. SHAP is an additive feature attribution framework that decomposes the model's prediction into the contributions of individual variables; the local prediction for borrower i is expressed as (Lundberg and Lee 2017)

$$g(z') = \varphi_0 + \sum_{j=1}^{M} \varphi_j \, z'_j, \qquad (4)$$

where φ_0 is the average predicted probability across the full sample (the base value) and φ_j is the SHAP value of variable j,[6] a weighted average of its marginal contributions across all possible orderings of variable entry. This allocation is the unique solution that simultaneously satisfies the three axioms of local accuracy, missingness, and consistency (Shapley 1953;

[6] The SHAP value, grounded in cooperative game theory, quantifies the contribution of each explanatory variable to the model's final prediction. Its starting point is the base risk score computed from the sample-wide average default rate alone, with no information about the borrower's consumption characteristics (the base value; for example, the average default rate of 3% converted into log-odds, the model's internal unit of computation, is -3.47). The SHAP value of each variable measures by how many points the borrower's particular consumption behavior (variable) raised ($+$) or lowered ($-$) the default risk score from this starting point. Adding all of a borrower's SHAP values to the base risk score reproduces the model's final prediction exactly, a property known as additivity.

Lundberg and Lee 2017). For tree-based models, we apply the TreeSHAP algorithm of Lundberg et al. (2020) to compute exact SHAP values efficiently across the full sample.

We interpret the fitted models using four diagnostics. First, variable importance at the full-sample level is measured by the mean absolute SHAP value: following Lundberg et al. (2020), we define the importance of each expenditure variable as

$$\text{Global Importance}_j = \frac{1}{n}\sum_{i=1}^{n}\left|\varphi_{ij}\right|, \qquad (5)$$

where n is the total number of borrowers and φ_{ij} is the SHAP value of variable j for borrower i; this identifies the spending-to-income categories that contribute most to discriminating credit risk. Second, partial dependence plots (PDPs), proposed by Friedman (2001), trace the pure marginal effect of a change in a given expenditure share on default probability while controlling for the influence of the other variables through marginalization, a mode of interpretation consistent with marginal analysis in traditional economics. Third, we verify monotonicity using surrogate beta coefficients and slope ratios: drawing on the local surrogate model framework of Ribeiro, Singh, and Guestrin (2016) (LIME, Local Interpretable Model-agnostic Explanations), we estimate local slopes over specific consumption ranges and test whether increases in an expenditure share translate consistently into increases or decreases in the default risk score, an essential step for confirming the practical stability of a credit scoring model. Fourth, variable-removal robustness checks re-estimate the model while cumulatively removing the main expenditure variables in order of their SHAP global importance and measure the mean and standard deviation of the resulting loss in discriminatory power (the AUROC drop) (Lundberg et al. 2020); this quantifies the extent to which each expenditure category contributes to preserving credit risk predictive power and also reveals whether the MyData variables are substitutable for one another.

3.2 Predictive performance

We split the full sample of 202,247 borrowers (default rate 3.1%) into a training sample of 161,797 borrowers and a test sample of 40,450 and compare the predictive performance of the baseline and disaggregated models. Table 6 presents the results.

[Table 6 about here.]

On the test sample, the baseline model attains an AUROC of 0.7957, indicating that financial and individual characteristic information alone identifies credit risk to a meaningful degree. Adding the disaggregated consumption behavior variables raises the AUROC of the all-variables model to 0.8136, an improvement of 0.0179 over the baseline. The significant-variables model, which ensures variable parsimony, maintains the same performance at 0.8136, showing that variable selection entails virtually no loss of information.

To evaluate predictive performance from multiple angles, we also analyze the Gini coefficient,[7] the Kolmogorov-Smirnov (K-S) statistic,[8] and the precision-recall area under the curve (PR-AUC).[9] On the test sample, the Gini coefficient of the disaggregated all-variables model is 0.6272, up 0.0358 from the baseline model (0.5914), and the K-S statistic likewise

---

[7] The Gini coefficient is computed from the area between the ROC curve and the diagonal (the random-classification line) and relates to the AUROC through Gini $= 2 \times \text{AUROC} - 1$. It ranges from 0 (no discriminatory power) to 1 (perfect discrimination) and serves as a summary measure of the overall discriminatory power of a credit risk model.

[8] The Kolmogorov-Smirnov (K-S) statistic measures the maximum distance between the cumulative distribution functions of the non-default and default groups, capturing discriminatory power in the threshold region where the two groups' score distributions diverge most. It ranges from 0 (no discrimination) to 1 (perfect separation), with larger values indicating that the model separates the two groups more effectively.

[9] The PR-AUC (Precision-Recall Area Under the Curve) is the area under the curve tracing the relation between precision and recall, and it is well suited to evaluating a model's predictive power in imbalanced-data environments with low default rates. It screens out the distorting effect of the majority non-default observations and focuses on how precisely the model identifies actual defaulted borrowers. It ranges from 0 to 1, with larger values indicating better identification of high-risk borrowers.

improves from 0.4627 to 0.4909 (+0.0282). The change in the PR-AUC is particularly noteworthy: in an imbalanced-data environment with a default rate of 3.1%, the PR-AUC excludes the distorting effect of the majority non-default observations and focuses on how precisely the model captures truly high-risk borrowers. The PR-AUC of the baseline model is 0.1150, rising to 0.1332 in the disaggregated all-variables model, indicating that combining consumption behavior information not only raises overall discriminatory power but also allows more precise identification of the high-risk borrowers that matter for credit risk management.

To assess generalizability, we examine the performance gap between the training and test samples. For the disaggregated all-variables model, the gap between the training-sample AUROC (0.8305) and the test-sample AUROC (0.8136) is only 0.0169; the Gini coefficient is 0.6610 in the training sample and 0.6272 in the test sample, the K-S statistic 0.5188 and 0.4909, respectively, and the PR-AUC 0.1475 and 0.1332, with the same pattern across the other metrics. These modest gaps suggest that the regularization parameters imposed at the model-specification stage operate effectively, limiting the scope for overfitting, and they support the stable generalizability of consumption-based credit risk prediction.

Finally, comparing the logit and XGBoost specifications for the baseline model alone, the AUROC of the XGBoost baseline model (0.7957) exceeds that of the logit baseline model (0.7551) by 0.0406. This suggests that nonlinear structure is embedded even in the baseline model composed only of financial and individual characteristic variables, and that after the consumption behavior variables are added, XGBoost continues to absorb residual nonlinearity that the linear model cannot capture.

To gauge the practical significance of the AUROC improvement, we convert the prediction results into borrower counts (Table 7). We rank the 40,450 borrowers in the test sample by predicted default probability and consider classifying the top 10%, or 4,045 borrowers, as high risk. The baseline model captures 488 actual defaulted borrowers in this segment, whereas the model augmented with the consumption behavior variables captures 533. Holding the number of classified borrowers fixed, the augmented model identifies 45 additional defaulted borrowers, and the share of actual defaulted borrowers among those classified rises from 12.06% to 13.18%.

[Table 7 about here.]

The contribution of the consumption behavior variables grows as the classification threshold tightens. When the top 5% are classified as high risk, the model identifies 42 additional defaulted borrowers, a 13.8% improvement over the baseline model, whereas at the top 30% the gain narrows to 34 borrowers, or 3.6%. Consumption behavior information thus contributes relatively more to identification in the high-risk segment, in line with the PR-AUC improvement documented above.

Discriminatory power also improves in the low-risk segment. Among borrowers in the bottom 10% of predicted risk, the actual default rate falls from 0.25% under the baseline model to 0.15% under the model with the consumption behavior variables. Consumption behavior information therefore sharpens not only the identification of high-risk borrowers but also the screening of creditworthy borrowers.

The results can also be interpreted holding the share of identified defaulted borrowers fixed. To identify 30% of defaulted borrowers, the baseline model must classify 2,751 borrowers as high risk, whereas the model with the consumption behavior variables requires only 2,234. This removes 517 non-default borrowers from the high-risk classification, an 18.8% reduction relative to the baseline model. At a 50% identification threshold, the number of classified borrowers likewise falls from 5,630 to 5,124, a decline of 506. The 0.0179 AUROC improvement corresponds to classification changes of this magnitude.

3.3 Variable importance

To identify which expenditure categories contribute materially to credit risk prediction, we examine the mean absolute SHAP value (Mean Abs SHAP) together with the decline in AUROC when a variable is removed (Mean AUROC Drop). The two measures are complementary: whereas Mean Abs SHAP captures the average magnitude of a variable's contribution to individual predictions, the AUROC Drop measures how much the model's discriminatory power deteriorates when the variable is permuted, thereby quantifying the variable's purely independent informational contribution net of its correlation structure with the other variables. Because absolute SHAP values are difficult to compare directly across models whose prediction units and scales differ, column 2 of Table 8 also reports each variable's share (in percent) of the total contribution of all variables to predictive variation, that is, the percentage of the model's overall evidence base that the variable accounts for when the model assesses default (Lundberg and Lee 2017; Lundberg et al. 2020).

[Table 8 about here.]

In the disaggregated model, the financial variable, the ratio of financial assets to income, records the highest importance of all variables, with a Mean Abs SHAP of 0.4085, a share of 28.49%, and an AUROC Drop of 0.0911. Log income follows with the second-highest contribution, a share of 19.38% (Mean Abs SHAP 0.2779, AUROC Drop 0.0402). That the two financial variables capturing income level and asset capacity occupy the top of the predictive-contribution ranking is consistent with the logit results, in which the marginal effects of these two variables dominated those of the consumption variables.

Among the consumption behavior variables, the most notable results concern telecommunications/income and convenience stores/income, which account for 11.25% and 10.09% of total predictive variation, respectively, the highest relative contributions among the consumption variables. Their AUROC Drops, 0.0196 and 0.0150, far exceed the AUROC Drop of the traditional credit risk indicator, the debt-to-financial-assets ratio (0.0052). Beyond a high Mean Abs SHAP ranking, this indicates that these variables carry independent predictive information that is not substituted by other variables in the model. Their strong contributions are consistent with the earlier finding that their associations became more pronounced once the nonlinearity of income was controlled for with income quintile dummies, and they stand out all the more in XGBoost, which learns nonlinear structures automatically.

Education/income records a Mean Abs SHAP of 0.0398 (share 2.77%) and an AUROC Drop of 0.0034, the highest contribution among the consumption behavior variables after telecommunications and convenience stores; just as education expenditure exhibited an independent negative association in the logit analysis, it carries independent predictive information in XGBoost that is not substituted by other variables. Apparel/income (Mean Abs SHAP 0.0361, AUROC Drop 0.0032), food delivery/income (0.0389, 0.0019), and taxis/income (0.0295, 0.0019) show contributions of a similar order. Medical and health/income ranks in the lower middle by Mean Abs SHAP (0.0208, share 1.45%), but its AUROC Drop (0.0028) exceeds those of sports and fitness (0.0010) and culture and leisure (0.0008), indicating relatively large independent information content among the risk-mitigating expenditures. Overall, several consumption behavior variables, including telecommunications, convenience stores, and education, contribute independently to default prediction alongside the financial variables; whether each raises or lowers risk is examined next.

The surrogate beta coefficients and segment-level slope analyses characterize the direction and shape of these effects (see Appendix Tables A4 and A5). The surrogate beta coefficient, estimated from a univariate OLS regression of each variable's SHAP values on its raw values, provides a linear approximation of the XGBoost nonlinear response function over the full sample. Convenience stores/income shows the largest positive surrogate beta (0.5117), and taxis/income (0.1734), telecommunications/income (0.1699), and food delivery/income (0.1549), categories sharing the attribute of recurring, routine consumption, are also positive, consistent with the significant positive logit estimates for convenience stores and telecommunications. In contrast, education/income (−0.0926), sports and fitness/income (−0.0899), apparel/income (−0.0821), and medical and health/income (−0.0243) carry negative surrogate betas, matching their significant negative logit associations, and the financial controls log income (−0.1584) and financial assets/income (−0.1559) are likewise negative. The segment-level slope analysis, based on PDPs, estimates the slope separately in the lower segment (at or below the 25th percentile of each variable's distribution) and the upper segment (at or above the 75th percentile) and measures the strength of any diminishing pattern with the slope ratio (high-segment slope divided by low-segment slope). A clear diminishing pattern emerges for the risk-increasing variables: for convenience stores/income the low-segment slope is steep at 7.8056, whereas the effect essentially vanishes in the upper segment at −0.0037 (slope ratio −0.0005); telecommunications/income falls from 3.2348 to 0.0267 (slope ratio 0.0083); taxis/income (slope ratio 0.0295) and food delivery/income (slope ratio 0.0200) show the same pattern. The structure is symmetric for the risk-mitigating variables: education/income declines from −3.3833 to −0.0331 (slope ratio 0.0098), sports and fitness/income from −1.6957 to −0.0218 (0.0129), and medical and health/income from −0.3965 to −0.0110 (0.0277), so the reduction in credit risk is concentrated in the range where such spending starts from even a small scale. Log income, by contrast, has a positive low-segment slope of 0.1289 that turns negative at −0.1770 in the upper segment (slope ratio −1.3728), a nonmonotonic reversal of direction consistent with the nonlinearity of the income-default relation documented in the income quintile analysis. This diminishing structure is not captured by the single coefficient of a logit model that assumes linearity, and it shows that the marginal effect of a given expenditure category on credit risk can differ with a borrower's existing level of spending.

4 Heterogeneity and Robustness

This section first examines heterogeneity in the predictive contribution of consumption behavior across income quintiles and age groups, which is a central part of our empirical analysis. We then assess the robustness of the estimates across estimation methods and measurement choices and discuss the limits of interpretation.

4.1 Subgroup analysis: income quintiles and age groups

To examine whether the incremental predictive power of consumption behavior varies across borrower groups, we re-estimate the models separately by income quintile and age group. For each subgroup, the baseline model includes the financial controls and gender but excludes consumption variables, and the increment is measured as the improvement in test-sample AUROC after adding them. Table 9 presents the results; subsample estimates are reported in Appendix Tables A6–A7.

[Table 9 about here.]

The incremental predictive power of consumption variables varies substantially across income quintiles. The full-sample increment of 0.0031 rises to 0.0318 in the second quintile (Q20–Q40) and 0.0206 in the third quintile, but remains close to the full-sample value in the lowest and highest quintiles (0.0029 and 0.0036). Thus, consumption information contributes most to discrimination among lower-middle-income borrowers. Within each income quintile, the narrower variation in income may reduce the discriminatory information contained in income, increasing the relative contribution of consumption variables. The estimates for the upper income quintiles are less precise because their test samples contain relatively few defaulted borrowers.

Because age does not vary within an age-group subsample, we exclude the age dummies and include only the financial controls and gender in the baseline model. The incremental AUROC improvement in the disaggregated model declines from 0.0033 for borrowers in their twenties or younger to 0.0009 for those in their fifties (Table 9, panel B). These point estimates suggest a broadly declining pattern through the fifties, although the age-specific increments are estimated less precisely than the income-group effects. The negative increment for borrowers aged 60 or above should be interpreted cautiously because the test sample contains only 105 defaults and exhibits an unusual reversal between training- and test-sample performance. Thus, the logit evidence on age heterogeneity is suggestive rather than statistically decisive.

The subgroup results reported thus far are based on the logit model. To examine whether the contribution of consumption variables also varies across subgroups in a nonlinear framework, we re-estimate the XGBoost model separately by income quintile and age group. We retain the hyperparameters selected using the full sample and do not retune them within each subgroup. The number of defaulted borrowers in the test samples ranges from 123 to 396 across income quintiles and from 105 to 432 across age groups.

[Table 10 about here.]

Across income quintiles, adding the consumption variables increases test-sample AUROC by 0.0179 in the full sample. The increase is largest in the second quintile (0.0443), followed by the third (0.0370), and ranges between 0.0170 and 0.0241 in the remaining quintiles. Importantly, this pattern closely mirrors the logit evidence: in both models, the incremental value of consumption is concentrated in the second and third income quintiles.

Across age groups, the AUROC increase is largest for borrowers aged 29 or below (0.0465). The increase is between 0.0146 and 0.0231 for borrowers aged 30 to 59 and only 0.0050 for those aged 60 or above. Thus, the incremental contribution of consumption information is greatest among younger borrowers and smallest among borrowers aged 60 or above.

Variable-level importance and permutation-based losses in predictive performance point in the same broad direction. For the seven consumption variables that are statistically significant in the risk-factor model (Table 4, column 5), their combined SHAP importance is highest in the second (35.57%) and third (35.29%) income quintiles. Across age groups, the combined share is highest for ages 30–39 (34.62%) and ages 29 or below (34.03%). When these variables are randomly permuted, the decline in test-sample AUROC is largest in the third income quintile (0.0500) and among borrowers aged 29 or below (0.0645). For borrowers aged 60 or above, the change is −0.0016, indicating no decline in predictive performance. Thus, both diagnostics indicate that the informational contribution of consumption is concentrated among younger borrowers, although the precise ranking of the two youngest groups differs across importance measures. The same qualitative pattern remains when all fifteen consumption variables are permuted.[10]

[Table 11 about here.]

[10] For the joint age-income distribution, see Appendix Table A8. Lower-income borrowers are concentrated among both younger and older age groups, so age and income do not fully overlap.

Taken together, the income- and age-based analyses show that the incremental predictive contribution of consumption information is particularly pronounced among lower-middle-income borrowers and younger borrowers.

We assess the statistical precision of the subgroup AUROC increments using 1,000 paired bootstrap replications (Efron 1979; Efron and Tibshirani 1993). In each replication, the baseline and consumption-augmented models are evaluated on the same resampled test observations, yielding confidence intervals and p-values for the incremental AUROC.

Bootstrap inference reinforces the income-group result. In the logit model, positive increments are statistically distinguishable from zero for the full sample and for the second and third income quintiles. By contrast, the age-specific logit increments are estimated less precisely: for borrowers aged 29 or below, the increment is 0.0033 ($p = .596$). We therefore interpret the logit evidence on age heterogeneity as suggestive rather than statistically decisive. Detailed bootstrap results for the logit model are reported in Appendix Table A9.

For XGBoost, the full-sample increment of 0.0179 is statistically significant, with a 95% confidence interval of [+0.0127, +0.0230] ($p = .001$). Across income quintiles, the first three quintiles and the fifth show statistically significant gains, while across age groups the three groups below age 50 do so. Detailed bootstrap results are reported in Appendix Table A10.

The bootstrap results show that the most robust cross-model evidence concerns lower-middle-income borrowers: the full sample and the second and third income quintiles are the only subsamples with statistically significant increments in both logit and XGBoost. The age evidence is more asymmetric across methods. The logit point estimates suggest greater value among younger borrowers but are imprecise, whereas XGBoost yields substantially larger and statistically significant gains for the younger groups.

4.2 Robustness of measurement and interpretation

We examine the robustness of the relation between MyData-based consumption behavior and credit risk using maximum-likelihood logit models and XGBoost interpreted through SHAP. The two approaches yield broadly consistent estimated association patterns. Financial assets-to-income and log income are the most important explanatory variables in both models. Among expenditure categories, telecommunications and convenience stores are positively associated with default risk, while education and medical and health spending are negatively associated. The signs of the logit coefficients are also consistent with those of the SHAP surrogate beta coefficients. Adding consumption variables improves test-sample AUROC by 0.0031 in the logit model and 0.0179 in XGBoost relative to the corresponding baseline models. These results suggest that the main association patterns and incremental predictive power of consumption behavior are not sensitive to the choice of estimation method.

The two approaches nevertheless provide complementary information. The logit model summarizes the average direction and magnitude of each association, whereas XGBoost-SHAP captures nonlinearities in the relationship between spending and default risk. Both risk-increasing expenditures, such as telecommunications and convenience stores, and risk-mitigating expenditures, such as education and medical and health spending, exhibit diminishing marginal effects, with marginal effects being largest at relatively low spending shares and declining thereafter. Similar nonlinear patterns emerge from both the surrogate beta coefficients and segment-level slope ratios, at the disaggregated level. The nonlinear relation between income and default risk is also robust. Specifications using income-quintile indicators produce a concave income-default relation while preserving the direction and significance of the main consumption variables (Table 5), and the same pattern emerges in the XGBoost partial-dependence analysis.

The incremental predictive contribution of consumption variables is further supported by the AUROC-drop analysis. Removing telecommunications/income and convenience-stores/income reduces test-sample AUROC by 0.0196 and 0.0150, respectively, compared with 0.0052 for the debt-to-financial-assets ratio. Thus, these consumption variables provide incremental predictive information beyond conventional financial indicators.

We also examine an alternative normalization that expresses each consumption category as a share of total expenditure rather than relative to income (Appendix Table A11). The principal signs and significance patterns are preserved in the disaggregated specification, with no variable that is significant in the main specification reversing sign. When total expenditure-to-income is additionally controlled for, the coefficient on total expenditure is insignificant and the consumption estimates and predictive performance remain essentially unchanged. Although the expenditure-share specification produces a somewhat higher AUROC, we retain the spending-to-income ratio as the main measure because it captures consumption burden relative to repayment capacity and is consistent with the denominator used for the financial controls.

Several caveats apply to the interpretation of these results. The estimated relations are conditional correlations rather than causal effects. The associations between particular expenditure categories and credit risk may reflect unobserved borrower characteristics related to repayment capacity, such as socioeconomic background, occupation, household composition, or financial literacy. We therefore interpret consumption variables as informative behavioral signals rather than causal determinants of default.

5 Conclusion

This paper examines whether granular consumption behavior contains information about consumer credit risk beyond conventional financial and individual characteristics and whether its informational value varies across borrowers. Using detailed financial transaction records collected through Korea's MyData regime, we relate disaggregated consumption expenditure to subsequent default using logit models and complementary XGBoost analysis.

We find that the composition of consumption contains predictive information beyond conventional financial indicators. In the logit analysis, seven expenditure categories remain significantly associated with default after financial controls are included: convenience stores, telecommunications, and cafés and snacks are positively associated with risk, whereas education, apparel, medical and health, and sports and fitness are negatively associated. Adding consumption variables also improves out-of-sample discrimination, increasing test-sample AUROC by 0.0031 in the logit model and by 0.0179 in XGBoost. The nonlinear analysis further identifies telecommunications and convenience-store spending as particularly informative predictors.

More importantly, the informational value of consumption varies systematically across borrowers. The logit analysis shows the strongest incremental predictive gains among lower-middle-income borrowers, a pattern reinforced by XGBoost and the accompanying importance measures. Consumption information also tends to be more informative among younger borrowers, although the age-specific logit estimates are less precise and the heterogeneity is more pronounced in the nonlinear model. These results suggest that the value of granular transaction information depends on the borrower's existing information environment rather than being uniform across the population. In particular, alternative behavioral information may be most useful where conventional financial indicators provide less complete discrimination of credit risk.

Our findings should be interpreted as conditional associations rather than causal effects. Consumption behavior may reflect unobserved borrower characteristics related to repayment capacity, and the analysis does not identify the mechanisms linking particular spending patterns to subsequent default. An additional limitation concerns sample selection. The final sample represents 20.2% of the original BankSalad users because borrowers without observed loans, users with zero recorded income, and observations with extreme values are excluded. Because some of these exclusions disproportionately affect higher-risk borrowers, the estimates may provide a conservative characterization of the relationship between consumption behavior and credit risk. The main results are nevertheless qualitatively robust to measuring category-level consumption as a share of total expenditure rather than income.

Future research using longer MyData panels could examine the dynamics of consumption behavior and credit risk and help clarify the mechanisms underlying the predictive relationships documented here.

References

Bakhtiari, A., S. M. Ali, B. P. S. Murthi, and E. M. Steffes. 2026. Prediction of consumer credit card risk from an analysis of spending categories: A hidden Markov model. *International Journal of Bank Marketing* 44:334–53.

Becker, G. S. 1962. Investment in human capital: A theoretical analysis. *Journal of Political Economy* 70:9–49.

Berg, T., V. Burg, A. Gombović, and M. Puri. 2020. On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies* 33:2845–97.

Bussmann, N., P. Giudici, D. Marinelli, and J. Papenbrock. 2021. Explainable machine learning in credit risk management. *Computational Economics* 57:203–16.

Chen, T., and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–94. ACM.

Chen, Y., R. Calabrese, and B. Martin-Barragan. 2024. Interpretable machine learning for unbalanced credit scoring datasets. *European Journal of Operational Research* 312:357–72.

D'Acunto, F., D. Hoang, M. Paloviita, and M. Weber. 2023. IQ, expectations, and choice. *Review of Economic Studies* 90:1152–84.

Di Maggio, M., and D. Ratnadiwakara. 2026. Invisible primes: Fintech lending with alternative data. *Management Science*, forthcoming.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7:1–26.

Efron, B., and R. J. Tibshirani. 1993. *An introduction to the bootstrap*. New York: Chapman and Hall.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29:1189–232.

Gathergood, J. 2012. Self-control, financial literacy and consumer over-indebtedness. *Journal of Economic Psychology* 33:590–602.

Gong, S., R. Levine, C. Lin, and W. Xie. 2024. Debtors at play: Gaming behavior and consumer credit risk. *Management Science* 70:5691–708.

Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80:223–55.

Iyer, R., A. I. Khwaja, E. F. P. Luttmer, and K. Shue. 2015. Screening peers softly: Inferring the quality of small borrowers. *Management Science* 62:1554–77.

Jagtiani, J., and C. Lemieux. 2019. The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management* 48:1009–29.

Karlan, D., and J. Zinman. 2009. Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica* 77:1993–2026.

Khandani, A. E., A. J. Kim, and A. W. Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34:2767–87.

Kim, H., H. Cho, and D. Ryu. 2018. An empirical study on credit card loan delinquency. *Economic Systems* 42:437–49.

Kuchler, T., and M. Pagel. 2021. Sticking to your plan: The role of present bias for credit card paydown. *Journal of Financial Economics* 139:359–88.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112:443–77.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2:56–67.

Lundberg, S. M., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30, 4765–74. Curran Associates.

Meier, S., and C. D. Sprenger. 2010. Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics* 2:193–210.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–44. ACM.

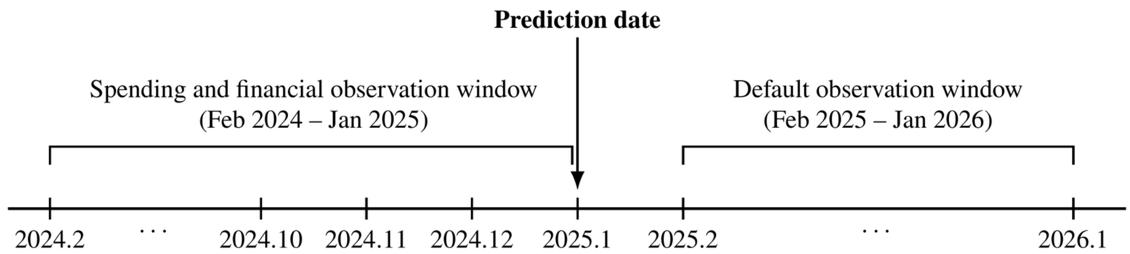
Shapley, L. S. 1953. A value for n-person games. In *Contributions to the Theory of Games*, Volume 2, 307–17. Princeton University Press.

Vissing-Jorgensen, A. 2021. Consumer credit: Learning your customer's default risk from what (s)he buys. Working Paper, University of California, Berkeley.

Zhu, N. 2011. Household consumption and personal bankruptcy. *Journal of Legal Studies* 40:1−37.

Figure 1

Timing structure of default prediction



Notes: This figure illustrates the forecasting horizon applied in the empirical analysis.

Table 1

Summary statistics: non-default versus default groups

	Non-default				Default			
	N	Mean	SD	Median	N	Mean	SD	Median
A. Individual characteristics								
Female (=1)	195,989	0.444	0.497	0.000	6,258	0.429	0.495	0.0
Age	195,989	37.698	9.940	35	6,258	39.342	11.969	37.0
B. Disaggregated expenditure categories								
Education/income	195,989	0.036	0.165	0.001	6,258	0.040	0.161	0.001
Culture and leisure/income	195,989	0.038	0.159	0.005	6,258	0.063	0.195	0.010
Sports and fitness/income	195,989	0.022	0.096	0.000	6,258	0.024	0.111	0.000
Beauty and personal care/income	195,989	0.040	0.180	0.004	6,258	0.054	0.230	0.006
Travel/income	195,989	0.071	0.292	0.005	6,258	0.066	0.278	0.003
Apparel/income	195,989	0.040	0.172	0.003	6,258	0.054	0.202	0.004
Medical and health/income	195,989	0.128	0.561	0.016	6,258	0.154	0.546	0.023
Taxis/income	195,989	0.029	0.121	0.003	6,258	0.051	0.142	0.007
Public transit/income	195,989	0.029	0.104	0.005	6,258	0.032	0.133	0.004
Childcare/income	195,989	0.004	0.033	0.000	6,258	0.005	0.030	0.000
Pets/income	195,989	0.007	0.050	0.000	6,258	0.010	0.047	0.000
Food delivery/income	195,989	0.044	0.158	0.005	6,258	0.074	0.181	0.013
Convenience stores/income	195,989	0.042	0.135	0.007	6,258	0.083	0.173	0.027
Housing/income	195,989	0.039	0.159	0.001	6,258	0.042	0.165	0.001
Telecommunications/income	195,989	0.090	0.455	0.007	6,258	0.263	0.673	0.032
Restaurants/income	195,989	0.148	0.493	0.043	6,258	0.226	0.611	0.070
Cafés and snacks/income	195,989	0.079	0.426	0.017	6,258	0.145	0.817	0.025
Bars and nightlife/income	195,989	0.017	0.075	0.001	6,258	0.027	0.091	0.002
Gifts and family occasions/income	195,989	0.008	0.055	0.000	6,258	0.011	0.059	0.000
C. Financial controls								
Income (10,000 KRW)	195,989	5,446.0	5,363.6	4,084.2	6,258	3,541.9	4,558.4	2,271.8
Financial assets/income	195,989	0.304	0.872	0.091	6,258	0.110	0.403	0.003
Debt/financial assets	195,989	337.2	1,039.4	17.4	6,258	1,134.4	1,815.9	197.5

Notes: This table reports the mean, standard deviation, and median of the analysis variables for the non-default group (N = 195,989) and the default group (N = 6,258, a default rate of 3.1%), measured over the financial observation window (February 2024 to January 2025). The unit of observation is the individual borrower. Consumption expenditure variables are spending-to-income ratios, defined as cumulative spending in each category over the window divided by cumulative income over the same period. *Female (=1)* is an indicator variable that equals one if the individual is female and zero otherwise. Income is reported in units of 10,000 KRW.

Table 2

Mean-difference tests between the non-default and default groups (t-tests)

Variable	Non-default	Default	t-statistic
A. Disaggregated expenditure categories			
Education/income	0.0358	0.0400	2.029**
Culture and leisure/income	0.0377	0.0634	10.351***
Sports and fitness/income	0.0217	0.0241	1.915*
Beauty and personal care/income	0.0400	0.0536	4.654***
Travel/income	0.0708	0.0663	−1.201
Apparel/income	0.0396	0.0535	5.402***
Medical and health/income	0.1281	0.1538	3.676***
Taxis/income	0.0292	0.0506	11.806***
Public transit/income	0.0293	0.0318	1.484
Childcare/income	0.0043	0.0048	1.191
Pets/income	0.0069	0.0099	4.953***
Food delivery/income	0.0443	0.0741	12.878***
Convenience stores/income	0.0421	0.0825	18.304***
Housing/income	0.0387	0.0418	1.541
Telecommunications/income	0.0895	0.2630	20.256***
Restaurants/income	0.1478	0.2258	10.010***
Cafés and snacks/income	0.0790	0.1452	6.384***
Bars and nightlife/income	0.0174	0.0268	8.061***
Gifts and family occasions/income	0.0085	0.0115	3.999***
B. Financial control variables			
Income (KRW 10,000)	5,446.0	3,541.9	−32.338***
Financial assets/income	0.3042	0.1098	−35.619***
Debt/financial assets	337.20	1,134.40	34.546***

Notes: This table reports tests of mean differences between the non-default and default groups for the consumption-behavior and financial variables. The full data period is February 2024 to January 2026 (financial variables observed February 2024 to January 2025; default observed February 2025 to January 2026), with 195,989 non-default borrowers and 6,258 defaulted borrowers (a default rate of 3.1%), for a total of N = 202,247. The Non-default and Default columns report each group's sample mean, and the t-statistic column reports the independent two-sample t-statistic. For each variable, we first apply Levene's test for equality of variances and use Student's t-test when equal variances are not rejected and Welch's t-test when the equal-variance assumption is rejected. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 3

Logit estimates: disaggregated model excluding financial controls

	Estimated coefficients			Marginal effect
	(1)	(2)	(3)	(4)
A. Expenditure-to-income ratios				
Education	−0.151*	−0.130		
	(0.086)	(0.086)		
Culture and leisure	0.169**	0.172**	0.170**	0.0050
	(0.068)	(0.067)	(0.066)	(0.0020)
Sports and fitness	−0.312**	−0.292**	−0.296**	−0.0088
	(0.150)	(0.149)	(0.146)	(0.0044)
Beauty and personal care	0.025	0.027		
	(0.072)	(0.072)		
Apparel	−0.130	−0.116		
	(0.082)	(0.081)		
Medical and health	−0.058**	−0.091***	−0.092***	−0.0027
	(0.026)	(0.028)	(0.027)	(0.0008)
Taxis	0.198**	0.234***	0.241***	0.0072
	(0.091)	(0.090)	(0.087)	(0.0026)
Pets	0.251	0.263		
	(0.230)	(0.230)		
Food delivery	0.117	0.192***	0.186**	0.0055
	(0.076)	(0.074)	(0.073)	(0.0022)
Convenience stores	0.672***	0.665***	0.680***	0.0202
	(0.077)	(0.077)	(0.073)	(0.0022)
Telecommunications	0.255***	0.247***	0.246***	0.0073
	(0.016)	(0.016)	(0.016)	(0.0005)
Restaurants	0.037	0.018		
	(0.023)	(0.024)		
Cafés and snacks	0.081***	0.076***	0.076***	0.0023
	(0.019)	(0.019)	(0.018)	(0.0006)
Bars and nightlife	−0.002	0.049		
	(0.169)	(0.167)		
Gifts and family occasions	0.101	0.092		
	(0.210)	(0.209)		

Notes: The table continues on the next page; complete notes follow the second part. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 3

Logit estimates: disaggregated model excluding financial controls (continued)

	Estimated coefficients			Marginal effect
	(1)	(2)	(3)	(4)
B. Individual characteristics				
Female		−0.021	−0.025	−0.0007
		(0.030)	(0.029)	(0.0009)
Age 30–39		−0.052	−0.051	−0.0015
		(0.039)	(0.039)	(0.0012)
Age 40–49		−0.056	−0.060	−0.0018
		(0.044)	(0.044)	(0.0013)
Age 50–59		0.275***	0.274***	0.0081
		(0.051)	(0.051)	(0.0015)
Age 60 or above		0.894***	0.897***	0.0267
		(0.063)	(0.063)	(0.0019)
Constant	−3.535***	−3.569***	−3.569***	
	(0.016)	(0.035)	(0.034)	
Observations (train)	161,797	161,797	161,797	
Observations (test)	40,450	40,450	40,450	
Pseudo-R^2	.0119	.0175	.0173	
Train AUROC	0.6744	0.6642	0.6649	
Test AUROC	0.6508	0.6591	0.6618	

Notes: This table reports logit maximum likelihood estimates of Equation (1) using the disaggregated spending-to-income ratio variables and excluding the financial controls. The sample covers the full data period from February 2024 to January 2026 and consists of 195,989 non-default borrowers and 6,258 defaulted borrowers (default rate 3.1%), for a total of N = 202,247 individuals; the sample is split 80:20 into training (161,797) and test (40,450) samples. Column 1 includes only the full set of consumption variables, column 2 adds the individual characteristics to column 1, and column 3 retains only the statistically significant consumption variables from column 2. Column 4 reports average marginal effects (AMEs) based on the column 3 estimates, expressing the average change in the default probability, in pp, when the corresponding expenditure share of income increases by 1 pp. Standard errors are in parentheses. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 4

Logit estimates: disaggregated model

	Estimated coefficients					Marginal effect
	(1)	(2)	(3)	(4)	(5)	(6)
A. Expenditure-to-income ratios						
Education		−0.289***		−0.297***	−0.307***	−0.0089
		(0.089)		(0.090)	(0.089)	(0.0026)
Culture and leisure		0.065		0.038		
		(0.070)		(0.072)		
Sports and fitness		−0.486***		−0.277*	−0.293*	−0.0085
		(0.152)		(0.151)	(0.150)	(0.0043)
Beauty and personal care		−0.042		−0.043		
		(0.073)		(0.074)		
Apparel		−0.211**		−0.184**	−0.194**	−0.0056
		(0.083)		(0.083)	(0.081)	(0.0024)
Medical and health		−0.169***		−0.152***	−0.160***	−0.0046
		(0.032)		(0.032)	(0.032)	(0.0009)
Taxis		0.103		0.135		
		(0.093)		(0.100)		
Pets		0.105		−0.023		
		(0.231)		(0.242)		
Food delivery		0.013		−0.036		
		(0.076)		(0.082)		
Convenience stores		0.364***		0.193**	0.187**	0.0054
		(0.081)		(0.090)	(0.081)	(0.0023)
Telecommunications		0.202***		0.155***	0.154***	0.0044
		(0.017)		(0.018)	(0.018)	(0.0005)
Restaurants		−0.060*		−0.041		
		(0.031)		(0.028)		
Cafés and snacks		0.053***		0.062***	0.057***	0.0017
		(0.019)		(0.020)	(0.020)	(0.0006)
Bars and nightlife		−0.149		0.013		
		(0.175)		(0.175)		
Gifts and family occasions		−0.110		−0.063		
		(0.226)		(0.230)		

Notes: The table continues on the next page; complete notes follow the second part. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 4

Logit estimates: disaggregated model (continued)

	Estimated coefficients					Marginal effect
	(1)	(2)	(3)	(4)	(5)	(6)
B. Financial controls						
Log income	−0.348***	−0.339***	−0.495***	−0.489***	−0.488***	−0.0141
	(0.012)	(0.015)	(0.013)	(0.016)	(0.016)	(0.0005)
Financial assets/income			−1.633***	−1.583***	−1.585***	−0.0458
			(0.075)	(0.075)	(0.075)	(0.0022)
Debt/financial assets			0.00023***	0.00023***	0.00023***	0.000007
			(0.00001)	(0.00001)	(0.00001)	(0.0000)
C. Individual characteristics						
Female	−0.087***	−0.059**	−0.081***	−0.056*	−0.055*	−0.0016
	(0.029)	(0.030)	(0.030)	(0.030)	(0.030)	(0.0009)
Age 30–39	0.075*	0.091**	0.154***	0.167***	0.161***	0.0047
	(0.039)	(0.039)	(0.040)	(0.040)	(0.040)	(0.0012)
Age 40–49	0.106**	0.133***	0.010	0.043	0.033	0.0010
	(0.044)	(0.045)	(0.045)	(0.046)	(0.045)	(0.0013)
Age 50–59	0.401***	0.442***	0.203***	0.246***	0.236***	0.0068
	(0.050)	(0.051)	(0.052)	(0.053)	(0.053)	(0.0015)
Age 60 or above	0.889***	0.935***	0.565***	0.612***	0.600***	0.0173
	(0.062)	(0.064)	(0.065)	(0.066)	(0.065)	(0.0019)
Constant	−0.805***	−0.897***	0.530***	0.458***	0.449***	
	(0.093)	(0.119)	(0.106)	(0.133)	(0.128)	
Observations (train)	161,797	161,797	161,797	161,797	161,797	
Observations (test)	40,450	40,450	40,450	40,450	40,450	
Pseudo-R^2	.0237	.0291	.0759	.0791	.0790	
Train AUROC	0.6590	0.6738	0.7581	0.7626	0.7625	
ΔTrain AUROC		+0.0148		+0.0045	+0.0044	
Test AUROC	0.6523	0.6664	0.7551	0.7582	0.7577	
ΔTest AUROC		+0.0141		+0.0031	+0.0026	

Notes: This table reports logit maximum likelihood estimates of the disaggregated consumption-variable model including the financial controls, together with average marginal effects (AMEs). The sample covers the full data period from February 2024 to January 2026 and consists of 195,989 non-default borrowers and 6,258 defaulted borrowers (default rate 3.1%), for a total of N = 202,247 individuals; the sample is split 80:20 into training (161,797) and test (40,450) samples. Column 1 is the income baseline model that controls only for log income and the individual characteristic variables, and column 2 adds the consumption variables to column 1. Column 3 is the financial baseline model that includes the full set of financial controls and the individual characteristic variables, the benchmark of the existing credit assessment framework; column 4 adds the consumption variables to column 3 (the full model); and column 5 is the risk-factor model that retains the significant variables. Column 6 reports AMEs based on the column 5 estimates: for the consumption variables, the change in the default probability (in pp) when the corresponding expenditure share of income increases by 1 pp; for log income, the change when income increases by 1%. The marginal effects for the female and age-group dummies are the change in the default probability associated with belonging to the category; in pp they equal the reported value multiplied by 100. ΔAUROC is the improvement of each model over its corresponding baseline: column 2 relative to column 1, and columns 4 and 5 relative to column 3. Blank cells indicate that the variable is not included in the model. Standard errors are in parentheses. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 5

Logit estimates: disaggregated model with income quintile dummies

	Estimated coefficients			Marginal effect
	(1)	(2)	(3)	(4)
A. Expenditure-to-income ratios				
Education		−0.291***	−0.288***	−0.0083
		(0.091)	(0.090)	(0.0026)
Culture and leisure		0.077		
		(0.071)		
Sports and fitness		−0.311**	−0.305**	−0.0088
		(0.152)	(0.151)	(0.0044)
Beauty and personal care		−0.027		
		(0.073)		
Apparel		−0.154*	−0.149*	−0.0043
		(0.084)	(0.082)	(0.0024)
Medical and health		−0.144***	−0.145***	−0.0042
		(0.032)	(0.032)	(0.0009)
Taxis		0.277***	0.300***	0.0086
		(0.100)	(0.095)	(0.0027)
Pets		−0.072		
		(0.244)		
Food delivery		0.044		
		(0.084)		
Convenience stores		0.420***	0.442***	0.0127
		(0.088)	(0.082)	(0.0024)
Telecommunications		0.166***	0.167***	0.0048
		(0.018)	(0.018)	(0.0005)
Restaurants		−0.012		
		(0.027)		
Cafés and snacks		0.072***	0.071***	0.0021
		(0.020)	(0.020)	(0.0006)
Bars and nightlife		0.085		
		(0.174)		
Gifts and family occasions		0.023		
		(0.226)		

Notes: The table continues on the next page; complete notes follow the second part. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 5

Logit estimates: disaggregated model with income quintile dummies (continued)

	Estimated coefficients			Marginal effect
	(1)	(2)	(3)	(4)
B. Income quintile dummies				
Income Q20–Q40 (low income)	−0.259***	−0.166***	−0.169***	−0.0049
	(0.037)	(0.040)	(0.040)	(0.0011)
Income Q40–Q60	−0.964***	−0.860***	−0.864***	−0.0249
	(0.044)	(0.048)	(0.048)	(0.0014)
Income Q60–Q80	−1.608***	−1.502***	−1.505***	−0.0434
	(0.055)	(0.058)	(0.058)	(0.0017)
Income Q80 and above (high income)	−1.766***	−1.660***	−1.664***	−0.048
	(0.057)	(0.061)	(0.060)	(0.0018)
C. Financial controls				
Financial assets/income	−1.656***	−1.597***	−1.597***	−0.046
	(0.077)	(0.076)	(0.076)	(0.0023)
Debt/financial assets	0.00022***	0.00022***	0.00022***	0.00001
	(0.00001)	(0.00001)	(0.00001)	(0.00000)
D. Individual characteristics				
Female	−0.136***	−0.104***	−0.107***	−0.0031
	(0.030)	(0.030)	(0.030)	(0.0009)
Age 30–39	0.221***	0.239***	0.237***	0.0068
	(0.040)	(0.040)	(0.040)	(0.0012)
Age 40–49	0.105**	0.146***	0.143***	0.0041
	(0.045)	(0.046)	(0.046)	(0.0013)
Age 50–59	0.279***	0.336***	0.330***	0.0095
	(0.052)	(0.053)	(0.053)	(0.0015)
Age 60 or above	0.589***	0.655***	0.649***	0.0187
	(0.064)	(0.066)	(0.066)	(0.0019)
Constant	−2.670***	−2.819***	−2.811***	
	(0.041)	(0.047)	(0.046)	
Observations (train)	161,797	161,797	161,797	
Observations (test)	40,450	40,450	40,450	
Pseudo-R^2	.088	.0925	.0924	
Train AUROC	0.7567	0.7629	0.7627	
ΔTrain AUROC		0.0062	0.006	
Test AUROC	0.7518	0.7565	0.7563	
ΔTest AUROC		0.0047	0.0045	

Notes: This table reports logit maximum likelihood estimates and average marginal effects (AMEs) for the disaggregated consumption-variable model including the financial controls, in which log income is replaced with income quintile dummies. The income quintile dummies take the lowest income quintile (bottom 20%) as the reference group. The sample covers the full data period from February 2024 to January 2026 and consists of 195,989 non-default borrowers and 6,258 defaulted borrowers (default rate 3.1%), for a total of N = 202,247 individuals; the sample is split 80:20 into training (161,797) and test (40,450) samples. Column 1 is the baseline model, column 2 is the full model, column 3 is the risk-factor model, and column 4 reports the marginal effects. The column 4 marginal effects are AMEs based on the column 3 estimates: for the consumption variables, the change in the default probability (in pp) when the corresponding expenditure share of income increases by 1 pp; for the income quintile dummies, the change relative to the reference group (the lowest income quintile). ΔTest AUROC is the improvement in the test-sample AUROC relative to the baseline model. Standard errors are in parentheses. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 6

Predictive performance across models

	Baseline model (Individual/financial)	Disaggregated model (2) All	Disaggregated model (3) Significant
A. Training sample (N = 161,797)			
AUROC	0.8039	0.8305	0.8267
		+0.0266	+0.0228
Gini coefficient	0.6078	0.6610	0.6534
		+0.0532	+0.0456
K-S statistic	0.4697	0.5188	0.5108
		+0.0491	+0.0411
PR-AUC	0.1144	0.1475	0.1429
		+0.0331	+0.0285
B. Test sample (N = 40,450)			
AUROC	0.7957	0.8136	0.8136
		+0.0179	+0.0179
Gini coefficient	0.5914	0.6272	0.6272
		+0.0358	+0.0358
K-S statistic	0.4627	0.4909	0.4930
		+0.0282	+0.0303
PR-AUC	0.1150	0.1332	0.1316
		+0.0182	+0.0166

Notes: This table reports the predictive performance of the baseline model and the disaggregated model using four metrics: AUROC, the Gini coefficient, the K-S statistic, and the PR-AUC. The full sample over the entire data period from February 2024 to January 2026 (N = 202,247; 195,989 non-default and 6,258 defaulted borrowers; default rate 3.1%) is split 80:20 into a training sample (N = 161,797) and a test sample (N = 40,450). The column labeled (2) All uses all expenditure variables and the column labeled (3) Significant uses only the significant expenditure variables. Improvements over the baseline model are reported in parentheses. Because the Gini coefficient satisfies Gini = 2 × AUROC - 1, the AUROC benchmarks of 0.60 and 0.70 presented earlier correspond to Gini coefficients of 0.20 and 0.40, respectively. A K-S statistic above 0.3 is conventionally regarded as acceptable and above 0.5 as good discriminatory power. Because the baseline level of the PR-AUC varies with the share of positive (default) observations, it is interpreted through the improvement over the baseline model rather than against a fixed threshold (a random model attains a PR-AUC of approximately 0.031). Detailed definitions of each metric are provided in the text footnotes.

Table 7

High-risk classification outcomes from adding consumption variables

	Classified borrowers	(1) Baseline model		(2) Disaggregated model		Difference
		Defaults	Precision	Defaults	Precision	
	(A)	(B)	(B/A)	(C)	(C/A)	(C) − (B)
Top 5%	2,023	304	15.03%	346	17.10%	+42
Top 10%	4,045	488	12.06%	533	13.18%	+45
Top 20%	8,090	780	9.64%	821	10.15%	+41
Top 30%	12,135	933	7.69%	967	7.97%	+34

Notes: This table reports, in terms of borrower counts, how adding the consumption behavior variables changes high-risk classification outcomes in the test sample (N = 40,450; 1,252 defaulted borrowers; default rate 3.1%). Borrowers are ranked from highest to lowest by the predicted default probability from the XGBoost model, and the top 5%, 10%, 20%, and 30% are classified as high risk. Model (1) is the baseline model that includes only the financial controls and individual characteristic variables, and model (2) is the full model that adds the 15 disaggregated consumption variables; they correspond to the baseline model and the disaggregated all-variables model of Table 6, respectively. The number of classified borrowers (A) is identical across the two models; only the composition of the borrowers classified as high risk differs. Defaults (B, C) are the numbers of borrowers classified as high risk who actually default during the default observation window. Precision is the share of actual defaulted borrowers among those classified as high risk (B/A, C/A); the default rate of the full test sample is 3.1%. The difference is the number of additional defaulted borrowers that model (2) identifies relative to model (1) (C − B). Because the test-sample default rate is adjusted through stratified sampling, the reported counts refer to that sample.

Table 8

SHAP variable importance: disaggregated model

	Mean Abs SHAP		Mean AUROC Drop	Std AUROC Drop
	(1)	(2)	(3)	(4)
Variable	Raw value	Share (%)		
Financial assets/income	0.4085	28.49	0.0911	0.0020
Log income	0.2779	19.38	0.0402	0.0009
Telecommunications/income***	0.1613	11.25	0.0196	0.0006
Convenience stores/income***	0.1447	10.09	0.0150	0.0008
Debt/financial assets	0.1398	9.75	0.0052	0.0006
Education/income***	0.0398	2.77	0.0034	0.0002
Food delivery/income	0.0389	2.71	0.0019	0.0003
Apparel/income***	0.0361	2.52	0.0032	0.0001
Taxis/income	0.0295	2.06	0.0019	0.0003
Medical and health/income***	0.0208	1.45	0.0028	0.0000
Cafés and snacks/income***	0.0177	1.23	0.0024	0.0001
Age 20s	0.0135	0.94	0.0011	0.0001
Age 60 or above	0.0131	0.91	0.0015	0.0001
Restaurants/income	0.0131	0.91	0.0010	0.0001
Age 40s	0.0123	0.86	0.0007	0.0001
Gifts and family occasions/income	0.0123	0.85	0.0008	0.0001
Sports and fitness/income***	0.0121	0.84	0.0010	0.0001
Pets/income	0.0117	0.82	0.0009	0.0001
Beauty and personal care/income	0.0094	0.66	0.0009	0.0000
Culture and leisure/income	0.0094	0.65	0.0008	0.0001
Age 30s	0.0048	0.34	0.0003	0.0000
Bars and nightlife/income	0.0043	0.30	0.0006	0.0000
Female	0.0023	0.16	0.0001	0.0000
Age 50s	0.0007	0.05	0.0001	0.0000

Notes: This table reports the predictive contributions of the disaggregated consumption behavior variables and the financial control variables in the XGBoost model. The sample covers the full data period from February 2024 to January 2026, with a training sample of N = 161,797 and a test sample of N = 40,450 (overall default rate 3.1%). Column 1 reports the raw values of Mean Abs SHAP, Mean AUROC Drop, and Std AUROC Drop, and column 2 reports each value's share of the sum across all variables (variable value / total sum × 100, in %). Mean Abs SHAP is the sample mean of the absolute SHAP values of each variable. Mean AUROC Drop is the average decline in AUROC under five-fold cross-validation when the variable is permuted, and it captures the amount of independent predictive information the variable carries. Std AUROC Drop is the standard deviation of the AUROC Drop across the five folds. Three asterisks after a variable name denote the consumption variables that are statistically significant in the corresponding logit model, and are shown for ease of reference; SHAP values themselves do not carry significance levels. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 9

Predictive contribution of consumption variables by subgroup: the logit model

	Default Rate	Test-sample Defaults	Baseline AUROC	Disaggregated model AUROC	Disaggregated model ΔAUROC
Full sample	3.09%	1,252	0.7551	0.7582	+0.0031
A. Income quintiles					
Q0–Q20	4.98%	382	0.7389	0.7418	+0.0029
Q20–Q40	4.99%	396	0.7373	0.7691	+0.0318
Q40–Q60	2.66%	225	0.7555	0.7761	+0.0206
Q60–Q80	1.47%	126	0.7998	0.8074	+0.0076
Q80–Q100	1.36%	123	0.8000	0.8036	+0.0036
B. Age groups					
Ages 29 or below	3.05%	266	0.7290	0.7323	+0.0033
Ages 30–39	2.78%	432	0.7793	0.7819	+0.0026
Ages 40–49	2.78%	281	0.7524	0.7544	+0.0020
Ages 50–59	3.83%	168	0.7131	0.7140	+0.0009
Ages 60 or above	7.11%	105	0.6944	0.6712	−0.0232

Notes: This table reports test-sample AUROCs of logit models estimated separately on income-quintile and age-group subsamples, together with the increments obtained by adding the fifteen disaggregated consumption variables. The baseline model excludes the consumption variables and includes financial assets/income, debt/financial assets, and the female dummy. The first row reports the full sample without subsample splits; panel A reports results by income quintile and panel B by age group. In panel A the baseline excludes log income, because income varies little within a quintile, and includes the age-group dummies; in panel B it includes log income and excludes the age-group dummies, because there is no within-group variation. The baseline AUROCs of the two panels are therefore not comparable with each other. ΔAUROC is the improvement in test-sample AUROC when the consumption variables are added. Subsample coefficient estimates are reported in Appendix Tables A6 and A7. Bootstrap confidence intervals and p-values for these increments are reported in Appendix Table A9.

Table 10

Predictive contribution of consumption variables by subgroup: the XGBoost model

	Default Rate	Test-sample Defaults	Baseline AUROC	Disaggregated model AUROC	Disaggregated model ΔAUROC
Full sample	3.09%	1,252	0.7957	0.8136	+0.0179
A. Income quintiles					
Q0–Q20	4.98%	382	0.7534	0.7704	+0.0170
Q20–Q40	4.99%	396	0.7414	0.7857	+0.0443
Q40–Q60	2.66%	225	0.7555	0.7925	+0.0370
Q60–Q80	1.47%	126	0.8054	0.8260	+0.0206
Q80–Q100	1.36%	123	0.8134	0.8375	+0.0241
B. Age groups					
Ages 29 or below	3.05%	266	0.7536	0.8001	+0.0465
Ages 30–39	2.78%	432	0.8192	0.8375	+0.0183
Ages 40–49	2.78%	281	0.7750	0.7981	+0.0231
Ages 50–59	3.83%	168	0.7440	0.7586	+0.0146
Ages 60 or above	7.11%	105	0.6986	0.7036	+0.0050

Notes: This table reports test-sample AUROCs of XGBoost models estimated separately on income-quintile and age-group subsamples, together with the increments obtained by adding the fifteen disaggregated consumption variables. The baseline model excludes the consumption variables and includes financial assets/income, debt/financial assets, and the female dummy. The first row reports the full sample without subsample splits; panel A reports results by income quintile and panel B by age group. In panel A the baseline excludes log income, because income varies little within a quintile, and includes the age-group dummies; in panel B it includes log income and excludes the age-group dummies, because there is no within-group variation. The baseline AUROCs of the two panels are therefore not comparable with each other. ΔAUROC is the improvement in test-sample AUROC when the consumption variables are added. Hyperparameters are held at the full-sample values reported in Appendix Table A3 and are not re-tuned by subgroup, so the subgroup increments are more likely to be understated than overstated. Bootstrap confidence intervals and p-values for these increments are reported in Appendix Table A10.

Table 11

Variable importance and predictive contribution of the significant consumption variables by subgroup: SHAP analysis of the XGBoost model

Variable	Q0–Q20	Q20–Q40	Q40–Q60	Q60–Q80	Q80–Q100
A. Income quintiles					
Mean Abs SHAP share (%)					
Telecommunications/income	9.04	13.37	12.88	10.10	9.82
Convenience stores/income	8.14	11.04	11.47	9.43	8.60
Education/income	4.34	1.60	2.59	1.69	5.37
Apparel/income	4.68	2.72	0.63	4.32	1.00
Medical and health/income	3.64	1.31	2.19	1.23	2.13
Cafés and snacks/income	1.28	3.59	4.48	2.49	2.30
Sports and fitness/income	1.29	1.94	1.05	1.18	3.25
Sum of the seven variables	**32.41**	**35.57**	**35.29**	**30.44**	**32.47**
Mean AUROC drop					
Seven significant variables	0.0340	0.0467	0.0500	0.0268	0.0351
All fifteen variables	0.0375	0.0518	0.0583	0.0332	0.0323
Variables	**Ages 29 or below**	**Ages 30–39**	**Ages 40–49**	**Ages 50–59**	**Ages 60 or above**
B. Age groups					
Mean Abs SHAP share (%)					
Telecommunications/income	14.58	10.54	9.62	3.64	8.74
Convenience stores/income	6.55	13.53	9.73	12.95	11.79
Education/income	3.65	1.22	2.34	1.85	0.56
Apparel/income	3.62	3.31	1.23	1.78	1.02
Medical and health/income	2.86	3.60	1.20	2.54	3.19
Cafés and snacks/income	1.83	1.28	1.26	4.17	3.90
Sports and fitness/income	0.94	1.14	0.71	2.56	1.60
Sum of the seven variables	**34.03**	**34.62**	**26.09**	**29.49**	**30.80**
Mean AUROC drop					
Seven significant variables	0.0645	0.0353	0.0257	0.0304	−0.0016
All fifteen variables	0.0878	0.0363	0.0311	0.0369	0.0044

Notes: This table reports the variable importance of the seven consumption variables that are statistically significant in the disaggregated logit model (Table 4, column 5), estimated separately on income-quintile and age-group subsamples of the XGBoost disaggregated model. Mean Abs SHAP share is the sample mean of the absolute SHAP values of a variable divided by the sum across all variables in the model, expressed in percent; the model includes all fifteen consumption variables, the financial control variables, and the individual characteristic variables, so the seven shares do not sum to 100. Mean AUROC drop is the average decline in the test-sample AUROC when the listed variables are randomly permuted, computed with five repetitions. The row "Seven significant variables" permutes the seven variables listed above and the row "All fifteen variables" permutes all fifteen consumption variables. Subsample controls follow the specification described in the notes to Table 10.

Appendix A. Additional Tables

Table A1

Sample construction and default rates of excluded borrowers

	N	Share of total	Defaulted borrowers	Default rate
A. Composition by reason for exclusion				
Total income = 0	169,419	16.90%	5,738	3.39%
Loan balance = 0, negative or unobserved	640,261	63.82%	6,057	0.95%
Loan balance = 0	514,900	51.32%	1,989	0.39%
Loan balance unobserved	125,339	12.49%	4,067	3.24%
Loan balance < 0	22	0.00%	1	4.55%
Deposit balance unobserved	125,339	12.49%	4,067	3.24%
Fund balance unobserved	18,150	1.81%	259	1.43%
ISA balance unobserved	9,075	0.90%	105	1.16%
Total expenditure = 0	1	0.00%	0	
Total excluded	669,947	66.78%	7,720	1.15%
B. Sample by preprocessing stage				
Non-default users at start of observation	1,003,293	100.00%	16,447	1.64%
Passing the variable-definition conditions	333,346	33.22%	8,727	2.62%
After first-stage preprocessing	333,327	33.22%	8,726	2.62%
After extreme-value trimming	266,661	26.58%	6,258	2.35%
Final analysis sample after stratified sampling	202,247	20.16%	6,258	3.09%

Notes: This table reports, for the 1,003,293 BankSalad users in non-default status at the start of the observation window, the borrowers excluded because the variables used in the analysis are not defined, together with the default rate of each group. Default status is measured over the default observation window (February 2025 to January 2026). Panel A reports the numbers of individuals and default rates by reason for exclusion, and panel B reports the evolution of the sample size across preprocessing stages. Because nonexistent records are recorded as missing, borrowers with a value of zero are those for whom the item is actually zero; a loan balance of zero thus means that the borrower holds no loans. Loan, deposit, and fund balances contain 22, 218, and 62 negative observations, respectively, which also fail the corresponding conditions and are excluded; panel A reports the negative cases separately for loan balances and, for deposit and fund balances, reports only the unobserved counts. Because a borrower can meet several exclusion criteria simultaneously, the reason-specific counts in panel A do not sum to the total excluded. The group with unobserved loan balances comprises the same borrowers as the group with unobserved deposit balances, namely users who did not provide asset information. The total excluded in panel A is computed on the six variable conditions shown in the panel. First-stage preprocessing removes, in addition to the panel A criteria, observations with missing individual characteristic variables such as age, which excludes a further 19 borrowers. Extreme-value trimming is applied variable by variable to the continuous variables, so a borrower is excluded even if only one variable takes an extreme value. In the stratified sampling stage, the default group is retained in full and only the non-default group is sampled, so that the default rate of the analysis sample is set to approximately 3.1%.

Table A2

Pearson correlation matrix of disaggregated expenditure variables

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22
Education	1.0	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.2	0.2	0.2	0.1	0.2	0.1	0.1	0.1	−0.2	0.1	0.0
Culture and leisure	0.1	1.0	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.0	0.1	0.3	0.3	0.2	0.2	0.3	0.2	0.3	0.1	−0.3	0.1	0.0
Sports and fitness	0.1	0.2	1.0	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.2	0.1	−0.2	0.1	0.0
Beauty and personal care	0.1	0.2	0.2	1.0	0.2	0.3	0.2	0.2	0.2	0.0	0.1	0.2	0.2	0.1	0.1	0.2	0.2	0.2	0.1	−0.2	0.1	0.0
Travel	0.1	0.2	0.2	0.2	1.0	0.2	0.2	0.2	0.3	0.0	0.1	0.2	0.2	0.2	0.1	0.3	0.2	0.2	0.1	−0.3	0.2	0.0
Apparel	0.1	0.2	0.2	0.3	0.2	1.0	0.2	0.2	0.3	0.0	0.1	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	−0.3	0.1	0.0
Medical and health	0.2	0.2	0.2	0.2	0.2	0.2	1.0	0.2	0.2	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.2	0.1	0.1	−0.3	0.1	0.0
Taxis	0.1	0.3	0.2	0.2	0.2	0.2	0.2	1.0	0.4	0.0	0.1	0.4	0.4	0.1	0.2	0.3	0.2	0.4	0.1	−0.3	0.2	0.0
Public transit	0.2	0.3	0.2	0.2	0.3	0.3	0.2	0.4	1.0	0.0	0.1	0.3	0.3	0.2	0.2	0.3	0.2	0.3	0.1	−0.4	0.2	0.0
Childcare	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	1.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	−0.1	0.0	0.0
Pets	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	1.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	−0.1	0.0	0.0
Food delivery	0.2	0.3	0.2	0.2	0.2	0.3	0.2	0.4	0.3	0.1	0.1	1.0	0.4	0.2	0.3	0.3	0.2	0.3	0.2	−0.4	0.2	0.0
Convenience stores	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.4	0.3	0.0	0.1	0.4	1.0	0.2	0.3	0.4	0.2	0.4	0.2	−0.4	0.2	0.1
Housing	0.2	0.2	0.1	0.1	0.2	0.2	0.3	0.1	0.2	0.1	0.1	0.2	0.2	1.0	0.2	0.3	0.2	0.1	0.1	−0.3	0.2	0.0
Telecommunications	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.0	0.1	0.3	0.3	0.2	1.0	0.2	0.1	0.1	0.1	−0.3	0.1	0.1
Restaurants	0.2	0.3	0.2	0.2	0.3	0.2	0.3	0.3	0.3	0.0	0.1	0.3	0.4	0.3	0.2	1.0	0.3	0.4	0.2	−0.4	0.2	0.0
Cafés and snacks	0.1	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.1	0.2	0.2	0.2	0.1	0.3	1.0	0.2	0.1	−0.2	0.1	0.0
Bars and nightlife	0.1	0.3	0.2	0.2	0.2	0.2	0.1	0.4	0.3	0.0	0.1	0.3	0.4	0.1	0.1	0.4	0.2	1.0	0.1	−0.3	0.2	0.0
Gifts and family occasions	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.1	1.0	−0.2	0.1	0.0
log(Income)	−0.2	−0.3	−0.2	−0.2	−0.3	−0.3	−0.3	−0.3	−0.4	−0.1	−0.1	−0.4	−0.4	−0.3	−0.3	−0.4	−0.2	−0.3	−0.2	1.0	−0.4	−0.1
Financial assets/income	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.2	0.0	0.0	0.2	0.2	0.2	0.1	0.2	0.1	0.2	0.1	−0.4	1.0	−0.1
Debt/financial assets	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	−0.1	−0.1	1.0

Notes: This table reports Pearson correlation coefficients for the 19 disaggregated expenditure variables and 3 financial control variables. The sample covers the full data period, February 2024 to January 2026 (N = 202,247, default rate 3.1%). All expenditure variables are measured as ratios of category spending to income. Variable codes: C1 = Education; C2 = Culture and leisure; C3 = Sports and fitness; C4 = Beauty and personal care; C5 = Travel; C6 = Apparel; C7 = Medical and health; C8 = Taxis; C9 = Public transit; C10 = Childcare; C11 = Pets; C12 = Food delivery; C13 = Convenience stores; C14 = Housing; C15 = Telecommunications; C16 = Restaurants; C17 = Cafés and snacks; C18 = Bars and nightlife; C19 = Gifts and family occasions; C20 = log(Income); C21 = Financial assets/income; C22 = Debt/financial assets. Correlation coefficients are rounded to one decimal place.

Table A3

XGBoost hyperparameter optimization results

Hyperparameter	Disaggregated model
n_estimators	100
learning_rate	0.021
max_depth	5
min_child_weight	2
subsample	0.60
colsample_bytree	0.58
lambda (λ)	2.19
alpha (α)	61.47
gamma (γ)	3.99
scale_pos_weight	17.06

Notes: This table reports the final hyperparameter values applied to the disaggregated model and summarizes the function and role of each parameter. Hyperparameters are optimized using the Optuna Bayesian optimization framework with stratified five-fold cross-validation.

Table A4

Surrogate beta coefficients: disaggregated model

	Surrogate beta coefficient	Direction
Variable	(1)	(2)
Convenience stores/income	0.5117	Positive (+)
Age 60 or above	0.1792	Positive (+)
Taxis/income	0.1734	Positive (+)
Telecommunications/income	0.1699	Positive (+)
Food delivery/income	0.1549	Positive (+)
Pets/income	0.1181	Positive (+)
Culture and leisure/income	0.0362	Positive (+)
Restaurants/income	0.0129	Positive (+)
Age 30s	0.0104	Positive (+)
Cafés and snacks/income	0.0089	Positive (+)
Female	0.0046	Positive (+)
Bars and nightlife/income	0.0033	Positive (+)
Debt/financial assets	0.0001	Positive (+)
Log income	−0.1584	Negative (−)
Financial assets/income	−0.1559	Negative (−)
Education/income	−0.0926	Negative (−)
Sports and fitness/income	−0.0899	Negative (−)
Apparel/income	−0.0821	Negative (−)
Age 20s	−0.0364	Negative (−)
Age 40s	−0.0339	Negative (−)
Medical and health/income	−0.0243	Negative (−)
Gifts and family occasions/income	−0.0239	Negative (−)
Beauty and personal care/income	−0.0190	Negative (−)
Age 50s	−0.0021	Negative (−)

Notes: This table reports the direction of the effect of the consumption behavior variables and the financial control variables on credit risk, measured by surrogate beta coefficients, for the disaggregated XGBoost model estimated on the full sample (N = 202,247, default rate 3.1%) over the full data period from February 2024 to January 2026, split 80:20 into a training sample (N = 161,797) and a test sample (N = 40,450). The surrogate beta coefficient is estimated from a univariate OLS regression of each variable's SHAP values on the raw values of that variable, and it provides a linear approximation of the XGBoost nonlinear response function over the full sample.

Table A5

Nonlinear marginal effects by segment: disaggregated model

	Low-segment slope	High-segment slope	Slope ratio (H/L)
Variable	(1)	(2)	(3)
Convenience stores/income	7.8056	−0.0037	−0.0005
Pets/income	5.8879	−0.0370	−0.0063
Telecommunications/income	3.2348	0.0267	0.0083
Taxis/income	2.0652	0.0609	0.0295
Food delivery/income	1.0647	0.0213	0.0200
Culture and leisure/income	0.1578	0.0131	0.0831
Log income	0.1289	−0.1770	−1.3728
Restaurants/income	0.1277	0.0007	0.0052
Debt/financial assets	0.0008	0.0000	0.0138
Gifts and family occasions/income	−4.6740	0.0023	−0.0005
Financial assets/income	−3.9568	−0.0187	0.0047
Education/income	−3.3833	−0.0331	0.0098
Apparel/income	−3.2600	−0.0118	0.0036
Sports and fitness/income	−1.6957	−0.0218	0.0129
Beauty and personal care/income	−0.7790	−0.0016	0.0021
Bars and nightlife/income	−0.7282	0.0091	−0.0125
Medical and health/income	−0.3965	−0.0110	0.0277
Cafés and snacks/income	−0.3407	−0.0004	0.0012

Notes: This table reports, for the disaggregated XGBoost model estimated on the full sample (N = 202,247, default rate 3.1%) over the full data period from February 2024 to January 2026, split 80:20 into a training sample (N = 161,797) and a test sample (N = 40,450), the slope of each variable in the lower segment (at or below the 25th percentile) and the upper segment (at or above the 75th percentile) of its distribution, computed from the partial dependence plot, together with the slope ratio (high-segment slope divided by low-segment slope). A slope ratio closer to zero in absolute value indicates that the marginal effect in the upper segment declines more strongly relative to the lower segment. Slopes and slope ratios are computed from unrounded values; a ratio recomputed from the rounded slopes reported in columns 1 and 2 may therefore differ slightly from column 3.

Table A6

Logit estimates by income quintile: disaggregated model

	(1) Q0–Q20	(2) Q20–Q40	(3) Q40–Q60	(4) Q60–Q80	(5) Q80–Q100
A. Individual expenditure ratio variables					
Education/income	−0.295***	−0.375*	−0.352	−0.926	0.371
	(0.105)	(0.225)	(0.442)	(0.906)	(0.923)
Culture and leisure/income	−0.052	0.393**	0.986***	1.701***	1.035
	(0.089)	(0.166)	(0.285)	(0.515)	(0.838)
Sports and fitness/income	−0.239	−0.589*	−1.001	−2.424	−8.998**
	(0.176)	(0.335)	(0.849)	(1.950)	(3.976)
Beauty and personal care/income	−0.054	0.174	−0.246	−0.569	0.775
	(0.089)	(0.138)	(0.357)	(0.773)	(0.941)
Apparel/income	−0.217**	−0.064	0.422	−0.364	0.019
	(0.102)	(0.184)	(0.328)	(0.906)	(1.351)
Medical and health/income	−0.126***	−0.143*	−0.153	−0.620	−1.251
	(0.036)	(0.083)	(0.205)	(0.474)	(0.864)
Taxis/income	0.215*	2.130***	4.694***	5.407***	3.156
	(0.111)	(0.362)	(0.798)	(1.679)	(2.298)
Pets/income	−0.434	0.015	1.206*	−1.246	2.006
	(0.328)	(0.513)	(0.661)	(2.408)	(1.553)
Food delivery/income	−0.094	0.842***	1.753***	1.136**	2.758
	(0.097)	(0.220)	(0.489)	(0.534)	(1.800)
Convenience stores/income	0.403***	1.786***	7.160***	4.651***	6.371***
	(0.095)	(0.330)	(1.038)	(1.571)	(1.977)
Telecommunications/income	0.113***	0.441***	0.875***	1.867***	2.631***
	(0.021)	(0.047)	(0.120)	(0.228)	(0.585)
Restaurants/income	−0.014	0.130*	0.257**	0.355	0.472
	(0.030)	(0.070)	(0.130)	(0.312)	(0.630)
Cafés and snacks/income	0.065***	0.142**	0.051	0.359	0.873*
	(0.021)	(0.062)	(0.225)	(0.293)	(0.469)
Bars and nightlife/income	−0.104	0.294	0.807	2.568***	2.412
	(0.211)	(0.403)	(0.801)	(0.784)	(1.701)
Gifts and family occasions/income	0.209	−0.088	−2.845	−0.702	−0.743
	(0.239)	(0.568)	(2.024)	(2.567)	(2.767)

Notes: The table continues on the next page; complete notes follow the second part. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table A6

Logit estimates by income quintile: disaggregated model (continued)

	(1) Q0–Q20	(2) Q20–Q40	(3) Q40–Q60	(4) Q60–Q80	(5) Q80–Q100
B. Financial control variables					
Financial assets/income	−0.992***	−3.247***	−7.032***	−11.846***	−17.139***
	(0.066)	(0.224)	(0.548)	(1.092)	(1.791)
Debt/financial assets	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
C. Individual characteristics					
Female	−0.145***	−0.199***	0.016	0.097	0.159
	(0.054)	(0.054)	(0.075)	(0.101)	(0.103)
Age 30–39	0.482***	0.341***	0.206**	0.151	−0.341**
	(0.068)	(0.068)	(0.095)	(0.158)	(0.169)
Age 40–49	0.337***	0.195**	0.159	0.028	−0.653***
	(0.078)	(0.080)	(0.112)	(0.166)	(0.173)
Age 50–59	0.448***	0.436***	0.106	0.283	−0.284
	(0.091)	(0.094)	(0.142)	(0.188)	(0.186)
Age 60 or above	0.411***	0.742***	0.900***	0.964***	0.730***
	(0.116)	(0.113)	(0.168)	(0.232)	(0.221)
Constant	−2.856***	−3.016***	−3.610***	−3.903***	−3.568***
	(0.061)	(0.066)	(0.102)	(0.159)	(0.165)
Observations (non-default)	38,434	38,431	39,373	39,853	39,898
Observations (default)	2,016	2,018	1,076	596	552
Pseudo-R^2	.0758	.0832	.1131	.1193	.1083
Train AUROC	0.7352	0.7463	0.7904	0.8121	0.7937
ΔTrain AUROC	+0.0106	+0.0237	+0.0228	+0.0202	+0.0117
Test AUROC	0.7418	0.7691	0.7761	0.8074	0.8036
ΔTest AUROC	+0.0029	+0.0318	+0.0206	+0.0076	+0.0036

Notes: This table reports logit maximum likelihood estimates of the disaggregated model with financial control variables, estimated separately on income-quintile subsamples. The sample covers the full data period, February 2024 to January 2026, with 195,989 non-default borrowers and 6,258 defaulted borrowers (default rate 3.1%; N = 202,247). Columns 1 to 5 report estimates for the subsamples from the lowest income quintile (Q0–Q20) through Q20–Q40, Q40–Q60, and Q60–Q80 to the highest quintile (Q80–Q100). We estimate within each quintile a model that excludes the expenditure-ratio variables and a model that includes them, and we report the difference in AUROC between the two models. Default rates by income quintile, from the lowest (Q0–Q20) upward, are 4.98%, 4.99%, 2.66%, 1.47%, and 1.36%: the bottom two quintiles are at similar levels, and above them default rates decline with income. Train and Test AUROC refer to the model that includes the consumption variables; ΔTrain and ΔTest AUROC are the improvements relative to the model that excludes them (financial and individual characteristic variables only) and measure the discriminatory contribution of the consumption variables in each quintile. Standard errors are in parentheses. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table A7

Logit estimates by age group: disaggregated model

	(1) Ages 29 or below	(2) Ages 30–39	(3) Ages 40–49	(4) Ages 50–59	(5) Ages 60 or above
A. Individual expenditure ratio variables					
Education/income	−0.617**	−0.271	−0.427***	−0.147	0.535
	(0.303)	(0.194)	(0.148)	(0.186)	(0.341)
Culture and leisure/income	−0.007	0.193	−0.112	0.155	−0.070
	(0.151)	(0.120)	(0.182)	(0.179)	(0.275)
Sports and fitness/income	−0.862**	−0.328	0.123	−0.457	−0.287
	(0.433)	(0.250)	(0.278)	(0.373)	(0.691)
Beauty and personal care/income	0.014	0.086	−0.127	−0.315	−0.037
	(0.160)	(0.118)	(0.162)	(0.221)	(0.284)
Apparel/income	0.010	−0.476***	−0.179	0.121	−0.260
	(0.148)	(0.169)	(0.217)	(0.182)	(0.304)
Medical and health/income	−0.202*	−0.094*	−0.237***	−0.102*	−0.129
	(0.115)	(0.055)	(0.076)	(0.059)	(0.086)
Taxis/income	0.325**	0.152	0.252	−0.171	0.356
	(0.148)	(0.187)	(0.311)	(0.427)	(0.581)
Pets/income	−0.244	−0.136	0.213	−0.220	0.548
	(0.539)	(0.436)	(0.492)	(0.673)	(0.845)
Food delivery/income	−0.054	−0.094	−0.055	−0.488	−0.080
	(0.161)	(0.134)	(0.188)	(0.388)	(0.475)
Convenience stores/income	0.153	0.248	0.317*	0.350	0.012
	(0.214)	(0.165)	(0.181)	(0.217)	(0.399)
Telecommunications/income	0.134***	0.215***	0.137***	0.122**	0.092
	(0.035)	(0.032)	(0.040)	(0.061)	(0.070)
Restaurants/income	−0.289**	−0.090	−0.081	0.051	0.046
	(0.133)	(0.064)	(0.065)	(0.039)	(0.070)
Cafés and snacks/income	0.089	0.043	0.076**	0.108**	−0.123
	(0.061)	(0.046)	(0.032)	(0.046)	(0.107)
Bars and nightlife/income	0.115	0.064	0.267	−0.730	2.208***
	(0.371)	(0.343)	(0.369)	(0.551)	(0.615)
Gifts and family occasions/income	0.520	−0.032	0.083	−0.613	−1.549
	(0.392)	(0.439)	(0.450)	(0.662)	(1.508)
B. Financial control variables					
Log income	−0.352***	−0.595***	−0.583***	−0.474***	−0.158***
	(0.036)	(0.029)	(0.033)	(0.041)	(0.057)
Financial assets/income	−1.507***	−1.464***	−1.611***	−2.129***	−1.866***
	(0.153)	(0.111)	(0.166)	(0.285)	(0.351)
Debt/financial assets	0.00033***	0.00029***	0.00018***	0.00018***	0.00006*
	(0.00002)	(0.00001)	(0.00002)	(0.00002)	(0.00003)
C. Individual characteristics					
Female	−0.126**	0.022	−0.097	−0.125	−0.063
	(0.063)	(0.052)	(0.066)	(0.085)	(0.114)
Constant	−0.543*	1.339***	1.332***	0.718**	−1.174**
	(0.287)	(0.243)	(0.281)	(0.350)	(0.481)
Observations (non-default)	45,434	75,950	47,880	20,567	6,158
Observations (default)	1,430	2,170	1,368	819	471
Pseudo-R^2	.0611	.0971	.0836	.0847	.0414
Train AUROC	0.7368	0.7893	0.7614	0.7548	0.6796
ΔTrain AUROC	+0.0084	+0.0068	+0.0041	+0.0051	+0.0198
Test AUROC	0.7323	0.7819	0.7544	0.7140	0.6712
ΔTest AUROC	+0.0033	+0.0026	+0.0020	+0.0009	−0.0232

Notes: This table reports logit maximum likelihood estimates of the disaggregated model with financial control variables, estimated separately on age-group subsamples. The sample covers the full data period, February 2024 to January 2026, with 195,989 non-default borrowers and 6,258 defaulted borrowers (default rate 3.1%; N = 202,247). Columns 1 to 5 report estimates for the subsamples of ages 29 or below, 30–39, 40–49, 50–59, and 60 or above. Because there is no within-group age variation in the age-group subsamples, the age-group dummies are excluded and the financial controls and the female dummy serve as control variables. Consumption variables are ratios of category spending to income, and log income is the log-transformed income. The age-group samples consist of 46,864 individuals aged 29 or below (1,430 defaults; 3.05%), 78,120 aged 30–39 (2,170; 2.78%), 49,248 aged 40–49 (1,368; 2.78%), 21,386 aged 50–59 (819; 3.83%), and 6,629 aged 60 or above (471; 7.11%). Within each age group we estimate a model that excludes the expenditure-ratio variables and a model that includes them, and we report the difference in the test-sample AUROC between the two models. Train and Test AUROC refer to the model that includes the consumption variables; ΔTrain and ΔTest AUROC are the improvements relative to the model that excludes them (financial and individual characteristic variables only) and measure the discriminatory contribution of the consumption variables in each age group. Standard errors are in parentheses. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table A8

Joint distribution of age groups and income quintiles

Age group	Q0–Q20	Q20–Q40	Q40–Q60	Q60–Q80	Q80–Q100	Default rate (%)
29 or below	30.4	28.0	21.7	12.4	7.5	3.05
30–39	15.9	18.3	23.4	23.1	19.4	2.78
40–49	15.8	15.6	15.7	23.4	29.5	2.78
50–59	19.1	17.3	15.1	19.2	29.3	3.83
60 or above	29.0	25.0	16.2	14.1	15.7	7.11

Notes: This table reports the joint distribution of age groups and income quintiles. Entries in the income-quintile columns are row percentages within each age group, as reported in Tables 9 and 10. The final column reports the default rate for each age group. The table shows that age and income are related but do not fully overlap. Among borrowers aged 29 or below, 58.4% are in the bottom two income quintiles; among borrowers aged 60 or above, the corresponding share is 54.0%. Conversely, borrowers aged 29 or below account for 35.2% of the bottom income quintile and 32.4% of the second income quintile. Thus, lower income is concentrated among both younger and older borrowers rather than being mechanically identified with youth. N = 202,247.

Table A9

Bootstrap inference for the AUROC increments: the logit model

	ΔAUROC	95% CI	p
Full sample	+0.0031	[+0.0008, +0.0053]	.004
A. Income quintiles			
Q0–Q20	+0.0029	[−0.0066, +0.0117]	.582
Q20–Q40	+0.0318	[+0.0184, +0.0450]	.001
Q40–Q60	+0.0206	[+0.0079, +0.0349]	.008
Q60–Q80	+0.0076	[−0.0059, +0.0210]	.304
Q80–Q100	+0.0036	[−0.0144, +0.0169]	.606
B. Age groups			
Ages 29 or below	+0.0033	[−0.0106, +0.0155]	.596
Ages 30–39	+0.0026	[−0.0019, +0.0058]	.210
Ages 40–49	+0.0020	[−0.0042, +0.0075]	.460
Ages 50–59	+0.0009	[−0.0075, +0.0080]	.820
Ages 60 or above	−0.0232	[−0.0515, −0.0016]	.038

Notes: This table reports bootstrap confidence intervals and p-values for the AUROC increments of the logit model reported in Table 9. Confidence intervals are percentile intervals from 1,000 paired bootstrap resamples of the test sample, in which the baseline and the consumption-augmented model are evaluated on the same resampled observations so that the increment is differenced within the resample. The two models are not re-estimated: the predicted probabilities for the test sample are held fixed and only the observations are resampled, and within each subgroup only that subgroup's test observations are resampled. The p-value is the two-sided proportion of resamples in which the increment is of the opposite sign, and its lower bound is 1/1,000 = .001. A 95% interval that excludes zero corresponds to a p-value below .05. The subsamples and the numbers of defaulted borrowers are identical to those of Table 9.

Table A10

Bootstrap inference for the AUROC increments: the XGBoost model

	ΔAUROC	**95% CI**	**p**
Full sample	+0.0179	[+0.0127, +0.0230]	.001
A. Income quintiles			
Q0–Q20	+0.0170	[+0.0057, +0.0277]	.001
Q20–Q40	+0.0443	[+0.0303, +0.0589]	.001
Q40–Q60	+0.0370	[+0.0180, +0.0576]	.001
Q60–Q80	+0.0206	[−0.0014, +0.0431]	.078
Q80–Q100	+0.0241	[+0.0073, +0.0424]	.004
B. Age groups			
Ages 29 or below	+0.0465	[+0.0304, +0.0631]	.001
Ages 30–39	+0.0183	[+0.0096, +0.0275]	.001
Ages 40–49	+0.0231	[+0.0106, +0.0363]	.001
Ages 50–59	+0.0146	[−0.0015, +0.0318]	.080
Ages 60 or above	+0.0050	[−0.0228, +0.0341]	.704

Notes: This table reports bootstrap confidence intervals and p-values for the AUROC increments of the XGBoost model reported in Table 10. Confidence intervals are percentile intervals from 1,000 paired bootstrap resamples of the test sample, in which the baseline and the consumption-augmented model are evaluated on the same resampled observations so that the increment is differenced within the resample. The two models are not re-estimated: the predicted probabilities for the test sample are held fixed and only the observations are resampled, and within each subgroup only that subgroup's test observations are resampled. The p-value is the two-sided proportion of resamples in which the increment is of the opposite sign, and its lower bound is 1/1,000 = .001. A 95% interval that excludes zero corresponds to a p-value below .05. The subsamples and the numbers of defaulted borrowers are identical to those of Tables 9 and 10.

Table A11

Logit estimates under alternative normalization: disaggregated model

	(1) Share of income	(2) Share of total expenditure	(3) (2) + total expenditure/income
A. Individual expenditure variables			
Education	−0.297***	−1.535***	−1.523***
	(0.090)	(0.426)	(0.426)
Culture and leisure	0.038	2.215***	2.230***
	(0.072)	(0.316)	(0.316)
Sports and fitness	−0.277*	−3.217***	−3.212***
	(0.151)	(0.675)	(0.676)
Beauty and personal care	−0.043	−1.077**	−1.072**
	(0.074)	(0.461)	(0.462)
Apparel	−0.184**	−0.538	−0.535
	(0.083)	(0.493)	(0.493)
Medical and health	−0.152***	−1.076***	−1.069***
	(0.032)	(0.213)	(0.213)
Taxis	0.135	3.004***	3.032***
	(0.100)	(0.368)	(0.368)
Pets	−0.023	0.543	0.554
	(0.242)	(0.809)	(0.808)
Food delivery	−0.036	2.114***	2.142***
	(0.082)	(0.345)	(0.345)
Convenience stores	0.193**	4.517***	4.570***
	(0.090)	(0.313)	(0.315)
Telecommunications	0.155***	3.082***	3.091***
	(0.018)	(0.115)	(0.116)
Restaurants	−0.041	−0.888***	−0.865***
	(0.028)	(0.232)	(0.233)
Cafés and snacks	0.062***	0.571**	0.572**
	(0.020)	(0.286)	(0.286)
Bars and nightlife	0.013	1.649***	1.665***
	(0.175)	(0.637)	(0.637)
Gifts and family occasions	−0.063	0.282	0.306
	(0.230)	(0.959)	(0.958)

Notes: The table continues on the next page; complete notes follow the second part. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table A11

Logit estimates under alternative normalization: disaggregated model (continued)

	(1) Share of income	(2) Share of total expenditure	(3) (2) + total expenditure/income
B. Financial control variables			
Log income	−0.489***	−0.409***	−0.398***
	(0.016)	(0.014)	(0.016)
Financial assets/income	−1.583***	−1.383***	−1.383***
	(0.075)	(0.071)	(0.071)
Debt/financial assets	0.00023***	0.00022***	0.00022***
	(0.00001)	(0.00001)	(0.00001)
Total expenditure/income			0.003
			(0.002)
C. Individual characteristics			
Female	−0.056*	0.030	0.031
	(0.030)	(0.032)	(0.032)
Age 30–39	0.167***	0.279***	0.275***
	(0.040)	(0.041)	(0.042)
Age 40–49	0.043	0.203***	0.198***
	(0.046)	(0.047)	(0.048)
Age 50–59	0.246***	0.449***	0.444***
	(0.053)	(0.055)	(0.055)
Age 60 or above	0.612***	0.881***	0.879***
	(0.066)	(0.068)	(0.068)
Constant	0.458***	−0.598***	−0.697***
	(0.133)	(0.117)	(0.142)
Observations (train)	161,797	161,797	161,797
Observations (test)	40,450	40,450	40,450
Pseudo-R^2	.0791	.1016	.1016
Train AUROC	0.7626	0.7797	0.7802
ΔTrain AUROC		+0.0171	+0.0176
Test AUROC	0.7582	0.7687	0.7692
ΔTest AUROC		+0.0105	+0.0110

Notes: This table reports logit maximum likelihood estimates of the disaggregated model re-estimated under an alternative normalization of the consumption variables. The sample covers the full data period, February 2024 to January 2026, with 195,989 non-default borrowers and 6,258 defaulted borrowers (default rate 3.1%; N = 202,247), split 80:20 into training (161,797) and test (40,450) samples. Column 1 is the specification in the text, with the consumption variables measured as ratios of category spending to income, and is identical to model (4) of Table 4. Column 2 measures the consumption variables as spending shares of total expenditure, and column 3 adds total expenditure/income as a control variable to column 2. The financial controls and individual characteristic variables are identical across the three columns. Because the two normalizations differ in scale, coefficient magnitudes are not directly comparable across columns; the comparison rests on the agreement of signs and significance. Standard errors are in parentheses. Significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.