https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–25
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Using Grocery Data for Credit Decisions

Jung Youn Lee,[a,*] Joonhyuk Yang,[b] Eric T. Anderson[c]

[a] Jones Graduate School of Business, Rice University, Houston, Texas 77005; [b] Mendoza College of Business, University of Notre Dame, Notre Dame, Indiana 46556; [c] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208

*Corresponding author

Contact: jungyoun.lee@rice.edu, https://orcid.org/0000-0002-6878-9832 (JYL); joonhyuk.yang@nd.edu, https://orcid.org/0000-0002-2846-5868 (JY); eric-anderson@kellogg.northwestern.edu, https://orcid.org/0000-0002-4981-7180 (ETA)

Received: August 4, 2022
Revised: June 5, 2023; November 1, 2023; March 10, 2024
Accepted: April 1, 2024
Published Online in Articles in Advance: July 15, 2024

https://doi.org/10.1287/mnsc.2022.02364

Abstract. Many consumers across the world struggle to gain access to credit because of their lack of credit scores. This paper explores the potential of a new alternative data source, grocery transaction data, for evaluating consumers' creditworthiness. Our analysis takes advantage of a unique, individual-level match of credit card data and supermarket loyalty card data. By developing credit scoring algorithms that either exclude or include grocery data, we illustrate both the incremental value of grocery data for credit decisions and its boundary conditions. We demonstrate that signals from grocery data can improve credit approval decisions, particularly for individuals who lack traditional credit scores. Furthermore, as a consumer establishes a relationship with lenders and builds a credit history, the marginal value of incorporating grocery data diminishes. These findings highlight the potential of grocery data in informing credit decisions and, consequently, in enabling financial institutions to extend credit to consumers who lack traditional credit scores.

History: Accepted by David Simchi-Levi, marketing.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2022.02364.

Keywords: grocery data • alternative data • consumer finance • credit card market • economics of data

1. Introduction

Recent advances in artificial intelligence and machine learning coupled with the evolution of large-scale data storage, access, and processing technologies have fueled interest among financial institutions in new data sources for credit scoring. Examples of these data sources include bill payment histories for phone, utility, and streaming services (McGurran 2023); transaction records from checking, savings, and money market accounts (FICO 2022); and rent payment histories (Fannie Mae 2021). These initiatives are driven by a dual objective: the pursuit of profit, such as the acquisition of new accounts, and the enhancement of social welfare, particularly by extending credit access to individuals who lack traditional credit scores. The absence of credit scores in financial markets, which often arises from limited or nonexistent credit histories, creates barriers for lenders to extend credit to these individuals, effectively excluding them from the formal financial systems. This challenge is not limited solely to individuals in developing and emerging economies (Demirgüç-Kunt et al. 2022), but extends to those in developed countries, including 45 million adults in the United States (Consumer Financial Protection Bureau 2016).

In this paper, we evaluate the potential of a new type of alternative data source, grocery transaction data, to assess consumers' creditworthiness. We begin by extracting signals of credit risk from grocery data and then examine the extent to which these signals predict consumer credit risk relative to traditional data available to lenders, such as income and credit scores. Through a simulation of hypothetical credit scoring and decision-making processes, we demonstrate that grocery data can offer informative signals of credit risk, leading to improved credit outcomes for creditworthy individuals and increased profitability for lenders. We also characterize the boundary conditions under which the use of grocery data adds no incremental value, which can shed light on when lenders might be incentivized to collect, acquire, and leverage alternative data. Specifically, we find that the incremental benefit of grocery data diminishes sharply as traditional credit scores or relationship-specific credit history becomes available. These findings highlight the potential for financial institutions to use grocery data to extend credit to individuals who lack traditional credit scores, while also demonstrating the limitations of this new data source.

Our empirical analysis is enabled by a novel proprietary data set from a multinational conglomerate operating in multiple cash-reliant, developing countries in Asia. We leverage data from the country in which the conglomerate is headquartered. The data sponsor owns a credit card issuer and a large-scale supermarket chain, which enables us to merge data from these two domains. In particular, using a customer identifier, we merge the supermarket's loyalty card data and the issuer's credit card spending and payment history at the individual level for the consumers who appear in both data sources between January 2017 and June 2019. The merged data allow us to observe how 30,089 consumers behave in the two seemingly different domains.

To assess the impact of incorporating grocery data into credit decisions, we take the perspective of a lender who faces two types of credit decisions. The first decision involves customer acquisition, in which the lender utilizes credit scoring algorithms to evaluate and screen new credit card applicants. The second decision concerns customer management, in which the lender's objective is to predict defaults among cardholders who have recently missed a credit card payment. For each of these scenarios, we build two sets of predictive algorithms: one that relies solely on traditional data, such as sociodemographic variables and traditional credit scores provided by credit bureaus, and another that incorporates grocery data to predict applicants' credit risk. The outcome variable for these algorithms is constructed based on consumers' credit card payment behaviors during the sample period, which is assumed to be unobserved by the lender at the time of prediction. We then simulate the lender's decision-making process using the predictions generated by each algorithm. This approach allows us to examine the incremental predictive power of grocery data compared with traditional data, holding other modeling choices fixed.

We begin our analyses by extracting signals of credit risk from grocery data. Our approach to feature engineering is motivated by our conversation with the manager of the data sponsor, who stated, "To work with these huge data sets, you need a strategy for summarizing the key pieces of data into meaningful variables. A naïve approach of simply throwing all our data at this problem without any structure is unlikely to work." This remark resonates with the comment made by a manager at one of the leading banks in the United States with whom we spoke. The manager mentioned that the major obstacle to utilizing large-scale, granular consumer data in the making of loans is not a lack of access to such data but rather a lack of knowledge on how to leverage them efficiently.

Our strategy for feature engineering is built on the premise that repeated behaviors in the grocery domain, which we refer to as grocery shopping habits, may contain signals of credit risk. Specifically, we posit that individuals who consistently demonstrate "good" behaviors in the grocery domain (e.g., purchasing healthy foods) are more likely to manifest good behaviors in financial domains (e.g., paying bills on time). This premise implies that an individual's grocery shopping habits can provide insights into the individual's financial behaviors and, consequently, credit risk. Guided by an extensive body of literature on habits, we construct variables that measure the level of consistency[1] or lack thereof in two broad dimensions of grocery shopping behaviors: what and how individuals buy. Grocery data lends itself particularly well to measuring general consumer traits, such as their tendency to engage in good behaviors, given that consumers make repeated and frequent choices within this domain.

We find that what one buys can explain what type of payer one is even after controlling for various sociodemographic variables and credit scores. For instance, buying cigarettes or energy drinks is associated with a higher likelihood of missing credit card payments or defaulting, whereas purchasing fresh milk or vinegar dressings is linked to consistently paying credit card bills on time. Using item-level survey ratings, we find suggestive evidence that buying healthier but less convenient food items is predictive of responsible payment behaviors. Furthermore, we observe a positive and robust correlation between displaying greater consistency in various dimensions of grocery shopping behavior and making timely credit card bill payments. For example, cardholders who consistently pay their bills on time are more likely to shop on the same day of the week, spend similar amounts across months, and purchase the same brands and product categories.

Next, we build credit scoring algorithms that either include or exclude the signals derived from grocery data. Our previous analysis of grocery shopping habits guides how we transform raw grocery data into inputs for these algorithms. We find that the incremental predictive gains largely depend on whether the lender knows the consumers' traditional credit scores from credit bureaus. In our final sample, credit scores are missing for roughly half of the consumers. Although we do not directly observe the source of missing data, evidence suggests that a significant number of consumers in developing and emerging economies, including the country represented in our data, lack credit scores as they do not have access to formal financial services and, therefore, generate no traditional financial data (Demirguc-Kunt et al. 2018). To allow for the potential differential impact of using grocery data for scored and unscored consumers, we perform analyses separately for each group.

To predict the credit risk of unscored consumers or those without credit scores, the lender often relies solely on sociodemographic variables, such as income. In these scenarios, incorporating grocery data significantly improves predictive accuracy, increasing out-of-sample predictive power by 3.11 to 7.66 percentage points, as measured by the area under the receiver operating characteristic curve (AUC). When it comes to consumers with credit scores, we find that grocery data, when used in isolation, can achieve predictive accuracy comparable to that of credit scores alone. This result implies that individuals' nonfinancial behaviors can provide credit risk signals of similar value to traditional credit scores. However, grocery data is not a perfect substitute for credit scores as there is a smaller yet positive incremental predictive gain from grocery data even relative to credit scores. More precisely, when both sociodemographic variables and credit scores are available, the incremental predictive power introduced by grocery data ranges from 0.359 to 2.51 percentage points in the out-of-sample AUC. Taken together, these results suggest that grocery data complements rather than substitutes traditional financial data, such as sociodemographic variables and credit scores.

To illustrate the impact of grocery data on customer acquisition decisions, we simulate the lender's decision of whether to approve a credit card applicant, assuming that the decision is based on the expected payoffs derived from acquiring the applicant. These expected payoffs are a function of the default predictions generated by the credit scoring algorithms. We worked with our data sponsor to develop and calibrate a model of credit extension decisions that accurately represents key aspects of the credit card approval process. Using this model, we simulate one potential approach for leveraging grocery data, in which the lender initially screens applicants using only standard data and subsequently incorporates grocery data as an additional screening device to refine the pool of initially approved applicants.

We find that implementing this two-stage decision rule leads to a 1.46% increase in per-person profits among applicants without credit scores. This increased profitability is driven by the improved risk profile of approved applicants as the rule effectively filters out defaulters, who experience a higher likelihood of rejection in the second stage than nondefaulters. By contrast, for applicants with credit scores, the impact on credit approval decisions and profitability is minimal with a 0.025% increase in per-person payoffs. These findings collectively suggest that, under the particular decision rule we consider, there may be a stronger motivation for the lender to acquire, collect, and leverage grocery data for evaluating applicants who lack a traditional credit score.

To further illustrate the value of grocery data in credit decisions, we consider an alternative decision: the management of existing cardholders who have recently missed a payment. In particular, we examine the extent to which incorporating grocery data contributes to predicting defaults among these cardholders. This predictive capability allows lenders to promptly implement targeted interventions, such as freezing the cards of cardholders with a sufficiently high default likelihood and simply sending payment reminders to those with a low default likelihood. Our findings indicate that grocery data can help better predict the default of a newly acquired cardholder who fails to make payments for the first credit card bills. However, beyond this case, the inclusion of grocery data offers limited additional value. Specifically, the incremental value of grocery data diminishes as cardholders establish and maintain their relationship with the issuer, generating a wealth of first party data, including detailed credit card spending data, which proves highly predictive of default. Overall, this observation establishes another boundary condition for the potential value of grocery data.

This paper contributes to the literature exploring the value of nontraditional, alternative data for credit scoring. Academics have examined various types of nontraditional data, including mobile phone usage data (San Pedro et al. 2015, Óskarsdóttir et al. 2019, Björkegren and Grissen 2020), text (Dorfleitner et al. 2016, Netzer et al. 2019), soft information (Iyer et al. 2016), social network (De Cnudde et al. 2019), digital footprints (Berg et al. 2020), verified consumer data (Chan et al. 2022), and education and employment history (Di Maggio et al. 2022). This paper assesses the information content of grocery data, which has received little attention in the domain of credit scoring. One exception is Vissing-Jorgensen (2021), who uses data from a Mexican retail chain that offers customers an option to buy products on credit to demonstrate the correlation between one's default risk and the items purchased at the retail store. Although similar in spirit, we explore a broader range of grocery shopping behaviors beyond what one buys. Further, to our knowledge, we are the first to explore the value of first party proprietary data as alternative data.

More broadly, this paper complements a new and rapidly growing literature on the economics of data. Although the perspective toward data as an input into a firm's production function is not new, the recent literature highlights the availability of large-scale data and the potential for data sharing across firms. Some studies in the stream of literature identify channels through which data and data sharing can generate welfare gains, such as prediction improvement and better customization (e.g., Bajari et al. 2019, Hughes-Cromwick and Coronado 2019, Jones and Tonetti 2020, Farboodi and Veldkamp 2021). This paper studies the context of first party data sharing across seemingly remote domains (i.e., marketing data from grocery retail and finance data from banks) and provides empirical evidence on the complementary nature of the data and resulting gains to firms. Further, it is often a central yet unverified premise of many theoretical models in the literature that improvement in prediction from data translates into an improvement in decision. We decompose the process of a firm's data usage into prediction and decision steps and demonstrate the effect of data on each step.

The rest of the paper proceeds as follows. Section 2 describes our empirical setting and data and defines the final sample for analysis. Section 3 explains how we extract relevant signals from grocery data, and Section 4 evaluates the predictive power of these signals. Section 5 illustrates the effect of incorporating grocery data into lenders' credit approval or customer acquisition decisions. Section 6 explores an alternative use case, examining the value of grocery data in managing existing customers. Section 7 discusses the managerial implications of our findings, and Section 8 concludes.

2. Data and Empirical Strategy

2.1. Data

We use a new, proprietary data set from an anonymous conglomerate that operates both a credit card issuer and a supermarket chain. The credit card issuer offers general-purpose credit cards that can be used at any merchant that accepts the associated processing network. The supermarket chain sells a wide range of products in various categories, including groceries, household supplies, clothing, and other general merchandise. Our analysis takes advantage of an individual-level match between credit card data and supermarket loyalty card data using a customer identifier. Observing first party data from both domains presents a unique opportunity, especially in the United States, where the banking and retail sectors are typically separate.[2]

2.1.1. Data from the Credit Card Issuer. We have four types of account-level data for approved cardholders: sociodemographic variables, credit scores, credit card spending data, and credit card payment history. The sociodemographic variables are self-reported and include monthly income, employment status, occupation, and number of dependents. The issuer supplements this information with credit scores purchased from a credit bureau. The issuer uses both sociodemographic variables and credit scores to evaluate credit card applications. We also have transaction-level data on cardholders' credit card spending between January 1, 2017, and December 31, 2018. For each transaction, we observe the date, transaction amount (both inflow and outflow), merchant category, and anonymized merchant name.

We observe cardholders' credit card payment history between June 14, 2017, and June 13, 2019. This data set includes the delinquency variable, which indicates the duration of any outstanding debt owed by the cardholder, as well as the payment variable, which records the monthly payment status (e.g., normal payment, no payment, overpayment).[3]

2.1.2. Data from the Supermarket. We have two types of data from the supermarket between January 1, 2017, and December 31, 2018: scanner panel data and store-level temporary price promotion data. The scanner panel data tracks purchases made with the supermarket's loyalty cards across different stores within the chain regardless of the payment method. This data includes the time stamp, item code, item hierarchy information, brand code, quantity purchased, price listed, price paid, customer identifier, transaction identifier, and store identifier. The item code is similar to the universal product code (UPC) used in the United States and is the most granular level of item definition in the data. The terms "item code" and "UPC" are used interchangeably hereafter. The item hierarchy information allows us to identify the nature of individual items. All UPCs are categorized into departments (e.g., consumer packaged goods), each of which is then broken down into sections (e.g., beverages). A section can be divided into product groups (e.g., mineral water), each of which can be further divided into product categories (e.g., flavored mineral water).

The store-level price promotion data records the following information for each UPC that was on sale at a specific store on a given date: store identifier, date and duration of in-store price promotions, regular and promoted prices, profit margin, on-shelf stock quantity, and quantity sold during the promotion. We merge the promotion data with the scanner panel data to assess whether a consumer purchased a sale item.

2.1.3. Merging Credit Card and Supermarket Data. By merging the credit card and supermarket data, we obtain a matched sample of 37,188 consumers. To ensure a one-to-one mapping between grocery shopping and credit card payment behaviors, we retain only consumers who satisfy two conditions: (1) the consumer is the only cardholder linked to the corresponding supermarket loyalty card, and (2) the consumer made purchases using only one credit card issued by our credit card issuer throughout the sample period. By doing so, we aim to mitigate concerns about the potential sharing of credit cards and/or supermarket loyalty cards among multiple individuals within a household.

Summary statistics for the full credit card sample, the full supermarket sample, and the matched sample are provided in Online Appendix A.1. Compared with the average credit card user, the average matched consumer has a lower monthly income, has a higher credit score, and spends more on the issuer's credit card. Compared with the average supermarket consumer, the average matched consumer spends more at the focal supermarket.

2.2. Empirical Strategy: Credit Approval Decisions

In this section, we primarily discuss the empirical strategy related to the first type of problem explored in this paper: the use of grocery data for a lender's credit approval decision or customer acquisition decision. Details on another problem we consider, which involves the use of grocery data for managing existing customers, can be found in Online Appendix D.

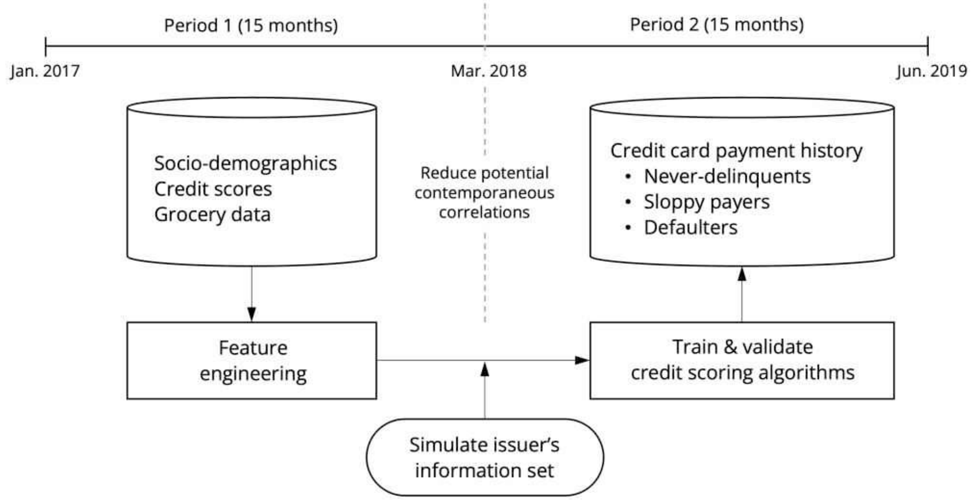
To assess the impact of incorporating grocery data into credit approval decisions, we take the perspective of a lender who evaluates and screens credit card applicants. We build a series of credit scoring algorithms, each of which assumes a different information set of the lender, and compare the resulting predictive accuracy. Specifically, our approach utilizes a two-period design, in which the sample period is split into two nonoverlapping periods, periods 1 and 2 in Figure 1. At the beginning of period 2, the lender is assumed to make approval decisions for the applicants based on consumer data available at that time. To simulate this information set, we use consumer data from period 1 (sociodemographic characteristics, credit scores, and grocery data) to create input features for the credit scoring algorithms.

The outcome variable in the credit scoring algorithms is constructed based on the consumers' credit card payment behaviors in period 2, which is assumed to be unobserved by the lender at the time of the approval decision. Specifically, following one of the segmentation schemes used by the data sponsor, we classify consumers into one of three credit card segments: never-delinquents (who always pay bills on time), sloppy payers (who miss payments periodically without defaulting), and defaulters.[4] In other words, we treat the lender's credit risk prediction problem as a classification task, in which the goal is to predict the applicants' credit card segments in period 2.

We operationalize the three credit card segments, guided by patterns observed in the data and discussions with company executives that provided the data, which broadly aligns with industry norms. In the credit card market, consumer default occurs in the form of consecutive delinquencies, and delinquency refers to a cardholder's failure to make at least the minimum payment by the end of a billing cycle. A cardholder is recorded as one-month delinquent as soon as the account is at least one day past due. If the cardholder fails to make a payment before the subsequent billing cycle (typically about a month apart from the current cycle), the cardholder is considered two-month delinquent. Depending on the issuer's policies, a cardholder who falls behind on payments for a certain number of consecutive months may be considered in default. In this paper, we define a consumer as a defaulter if the consumer entered a two-month delinquency at any point during period 2.[5] A consumer is classified as a sloppy payer if the consumer entered a one-month delinquency at least once during period 2 but never entered a two-month delinquency. A consumer is categorized as never-delinquent if the consumer never missed a payment throughout period 2.

Our empirical design offers two advantages. First, it allows us to assess the marginal impact of incorporating grocery data into credit decisions, which is the main goal of this paper. Specifically, because grocery data is available in period 1, we can construct credit scoring algorithms that either include or exclude grocery data and compare the resulting credit decisions.

Figure 1. Data and Empirical Strategy



Second, our design reduces the potential risk of falsely attributing the indicators of temporal financial shocks to grocery shopping habits (equivalent to individual fixed effects) as we use lagged grocery shopping behaviors in period 1 to predict credit risk in period 2.

2.3. Sampling and Summary Statistics

Among the matched consumers, we further narrow our focus to consumers who satisfy three conditions. First, they made at least five shopping trips to the supermarket during period 1, so we have enough observations to characterize their grocery shopping habits. Second, they did not default according to our definition of default during period 1. This condition is to capture individuals' grocery shopping behaviors independent of any temporal shocks that could have led them to default. Third, conditional on not entering a two-month delinquency during period 2, they did not enter a one-month delinquency in the last month of period 2. As we do not observe whether the last month delinquency in the data led to a two-month delinquency, we cannot conclude whether they would have been categorized as sloppy payers or defaulters.

Applying the three filters leaves us with 30,089 consumers. Among the sample consumers are 24,315 never-delinquent payers (81%), 3,576 sloppy payers (12%), and 2,198 defaulters (7%). Table 1 presents summary statistics for selected variables across the three segments. The three segments significantly differ based on key observable characteristics commonly used to evaluate credit risk, such as income and credit scores. Specifically, the average sloppy payer has the highest income, followed by never-delinquents and then defaulters. Regarding credit scores, the data sponsor's market uses a scale ranging from 300 to 900 with a score above 700 typically indicative of financial health. The average never-delinquent payer has the highest credit score, and the average sloppy payer has a credit score around the threshold of being considered financially healthy. The relatively high average credit scores across all segments may follow from the fact that all sample consumers have been approved for credit.[6] The average sloppy payers missed payments 2.58 times in period 2, whereas defaulters missed 1.48 times (excluding the delinquency that eventually led to default).

We note that credit scores are missing for 49.7% of the consumers in our final sample (14,952 out of 30,089 consumers). In our empirical setting, the absence of credit scores can be attributed to two reasons: either the credit card issuer did not purchase credit scores for those consumers or the consumers themselves do not have credit scores. Although we do not directly observe the source of missing data, there is ample evidence that a significant portion of consumers in developing and emerging markets, including the country represented in our data, are unbanked or have limited or no access to formal financial services, which makes it difficult for them to establish a credit history and obtain credit scores (e.g., Demirgüç-Kunt et al. 2022).

Table 2 compares consumers with credit scores ("scored" consumers) and those without credit scores ("unscored" consumers) based on selected variables. We find that unscored consumers tend to be riskier in terms of both ex ante and ex post credit risk. Specifically, the average unscored consumer has a lower income, which is typically associated with higher credit risk. Further, once approved, the average unscored consumer missed payments or defaulted on the focal credit card more than the average scored consumer. To allow for the potentially differential impact of incorporating grocery data for these two groups, we build credit scoring algorithms and simulate credit extension

Table 1. Consumer Characteristics by Segment

	Never-delinquents (N)	Sloppy payers (S)	Defaulters (D)	Statistical differences[a]
Monthly income, US$	12,964	13,217	10,704	*N* = *S* > *D*
	(19,851)	(22,092)	(19,106)	
Credit score	719	696	675	*N* > *S* > *D*
	(59.5)	(72.2)	(85.4)	
Monthly credit card spend, US$	2,199	1,494	1,596	*N* > *S* = *D*
	(3,374)	(2,523)	(3,159)	
Monthly supermarket spend, US$	491	459	376	*N* > *S* > *D*
	(426)	(410)	(370)	
Number of one-month delinquencies	–	2.58	1.48	*S* > *D*
	–	(2.82)	(2.36)	
Number of consumers	24,315	3,576	2,198	

Notes. The table reports the mean and standard deviation (in parentheses) of the variables in each sample. Monthly credit card and supermarket spending are based on data between January 2017 and March 2018 (period 1). The number of one-month delinquencies is based on data between April 2018 and June 2019 (period 2). As credit scores are missing for some credit card customers (see Table 2 for details), the reported statistics are conditional on nonmissing values.

[a]We perform both one-way ANOVA and pairwise *t*-tests to test for a difference in the mean of the focal variable between the three segments. Reported differences are based on the 95% confidence level.

Table 2. Summary Statistics for Unscored and Scored Consumers

	Unscored consumers	Scored consumers	Statistical differences[a]
Monthly income, US$	10,871 (17,530)	14,763 (22,156)	<0.001
Monthly credit card spending, US$	2,509 (3,643)	1,638 (2,808)	<0.001
Monthly supermarket spending, US$	488 (431)	469 (411)	<0.001
Credit card payment behavior in period 2			
Pr(Never-delinquent)	0.772 (0.420)	0.844 (0.363)	<0.001
Pr(Sloppy payer)	0.136 (0.343)	0.102 (0.303)	<0.001
Pr(Defaulter)	0.092 (0.290)	0.054 (0.226)	<0.001
Number of consumers	14,952	15,137	

[a] We report the *p*-values from two-sided *t*-tests to test for a difference in the mean of the focal variable between the two groups.

decisions separately for each group in the subsequent sections.

3. Extracting Signals from Grocery Data

This section illustrates how we extract signals of credit risk from grocery data and demonstrates how and to what extent the derived signals are correlated with credit risk. Our approach to feature engineering is based on the premise that individuals who consistently demonstrate good behaviors in the grocery domain are more likely to manifest good behaviors in financial domains. With this premise in mind, we explore two broad dimensions of grocery shopping habits: what and how one buys in a grocery store. We explain in detail how we operationalize each dimension.

3.1. Signals in Grocery Data: What to Buy

Among many product categories in the supermarket data, we focus on food items as the consumption of these items is often associated with various consumer habits (Khare and Inman 2006, Verhoeven et al. 2012). Assuming that food expenditures can serve as a valid proxy for food consumption, we explore how food expenditure is correlated with credit card payment behaviors.[7] Specifically, we construct two types of purchase indexes. First, we create a binary variable for each food item that indicates whether a consumer ever purchased the item in period 1 (*whether to buy* variable). Second, we pool a consumer's food item purchases in period 1 and compute the share of expenditure allocated to each item (*expenditure share* variable).

We then estimate the following multinomial logit model:

$$\Pr(Y_i = j \mid G_i, \boldsymbol{X_i}) = \frac{\exp(\alpha_j + G_i\beta_j + \boldsymbol{X_i'}\gamma_j)}{\sum_{k\in\{n,s,d\}}\exp(\alpha_k + G_i\beta_k + \boldsymbol{X_i'}\gamma_k)}. \tag{1}$$

The dependent variable is the probability that individual i belongs to a particular segment j in period 2: a never-delinquent ($j = n$), a sloppy payer ($j = s$), or a defaulter ($j = d$). Grocery shopping behaviors in period 1 are captured by G_i, and β_j is the parameter of interest. The vector $\boldsymbol{X_i}$ includes controls for sociodemographic characteristics (such as the number of dependents, monthly income, employment status, and occupation), credit scores, and the log of total grocery expenditure. We estimate the regression separately for each food item.[8]

Table 3 presents the top 10 items for each segment in order of the magnitude of the average marginal effects. Panels A and B show the results when *whether to buy* and *expenditure share* variables are used as the focal grocery variables, respectively. In panel A, fresh milk is ranked first on the list for never-delinquents, which indicates that purchasing fresh milk, relative to purchasing any other item, leads to the greatest increase in the probability of being a never-delinquent. In panel B, we find that increasing the share of grocery expenditure allocated to vinegar dressings by one standard deviation has the largest impact on increasing the probability of being a never-delinquent.

From a visual inspection of the items in Table 3, the three segments exhibit differences along at least two dimensions. First, they appear to buy items of varying levels of healthiness. In panel A, never-delinquents are more likely to buy healthy items, such as fresh milk, pulses (beans), fresh yogurt, and fruits and vegetables. Purchase of items with lower healthiness, such as cigarettes, energy drinks, and canned meat, is associated with a greater probability of being a defaulter. Further, both sloppy payers and defaulters appear to spend a larger share of their budget on fish and meat products, such as mutton and offal.

Second, never-delinquents tend to allocate a greater share of their grocery expenditure to less convenient items or items that require more time to transform purchases into consumption. In panel B, we find that they spend more on pantry staples, such as frozen cooking

Table 3. Food Items Purchased by Credit Card Segment

	Never-delinquents	Sloppy payers	Defaulters
	Panel A. G_i: Whether purchased an item		
1	Fresh milk	Cooked fish	Cigarettes
2	Dry bread	Deli pasta	Energy drinks
3	Ready-to-eat deli product	Imported snacks: B[a]	Canned meat
4	Imported snacks: A[a]	Deli seafood	Shellfish and mollusks
5	Pulses (beans)	Shellfish and mollusks	Beef
6	Flours	Deli meat	Offal
7	Fresh yogurt	Canned meat	Mortadella
8	Fruits and vegetables	Deli fish	Canned fish
9	Bulk ice cream	Imported snacks: C[a]	Chicken
10	Biscuits	Deli sausages	Deli meat
	Panel B. G_i: Expenditure share of an item		
1	Vinegar dressing	Mutton	Deli seafood
2	Ready-to-eat deli product	Deli seafood	Cooked fish
3	Frozen cooking fat	Cooked fish	Deli sausages
4	Imported snacks: A[a]	Deli pasta	Sliced cheese
5	Flours	Sliced cheese	Mortadella
6	Pulses (beans)	Imported snacks: B[a]	Fresh white bread
7	Fresh yogurt	Chilled cakes	Deli sliced meat
8	Fresh milk	Deli sausages	White cheese
9	Frozen pastries	Deli sliced meat	Imported snacks: D[a]
10	Cereals	White cheese	Deli meat

Notes. We run 162 separate regressions in Equation (1), each using one of the 162 food items as the focal grocery variable G_i. The items are listed in order of magnitude of the average marginal effects for each segment.

[a] Exporting countries are deidentified.

fat and flour, which are typically consumed not on their own but used as inputs for cooking. Sloppy payers and defaulters appear to purchase less time-intensive and prepared food items, such as deli seafood and deli pasta.

Although the visual inspection of individual grocery items provides suggestive evidence that buying healthier or less convenient items is associated with paying credit card bills on time, the analysis has some limitations. In particular, it does not speak to the relationship between credit risk and one's overall purchase profile and does not take into account the potential interplay between healthiness and convenience. Further, the analysis relies on intuitive but ad hoc measures of healthiness and convenience. In Section 3.3, we conduct a more systematic investigation of the link between credit risk and perceived healthiness and convenience of a grocery basket as measured by survey questions.

3.2. Signals in Grocery Data: How to Buy

Recognizing that there are countless ways to characterize how individuals buy, we narrow attention to the measures of consistency in behaviors to capture habits. In particular, we explore the consistency in four types of behaviors: consistency in (1) the timing of shopping, (2) spending amounts, (3) items purchased, and (4) taking advantage of promotions.

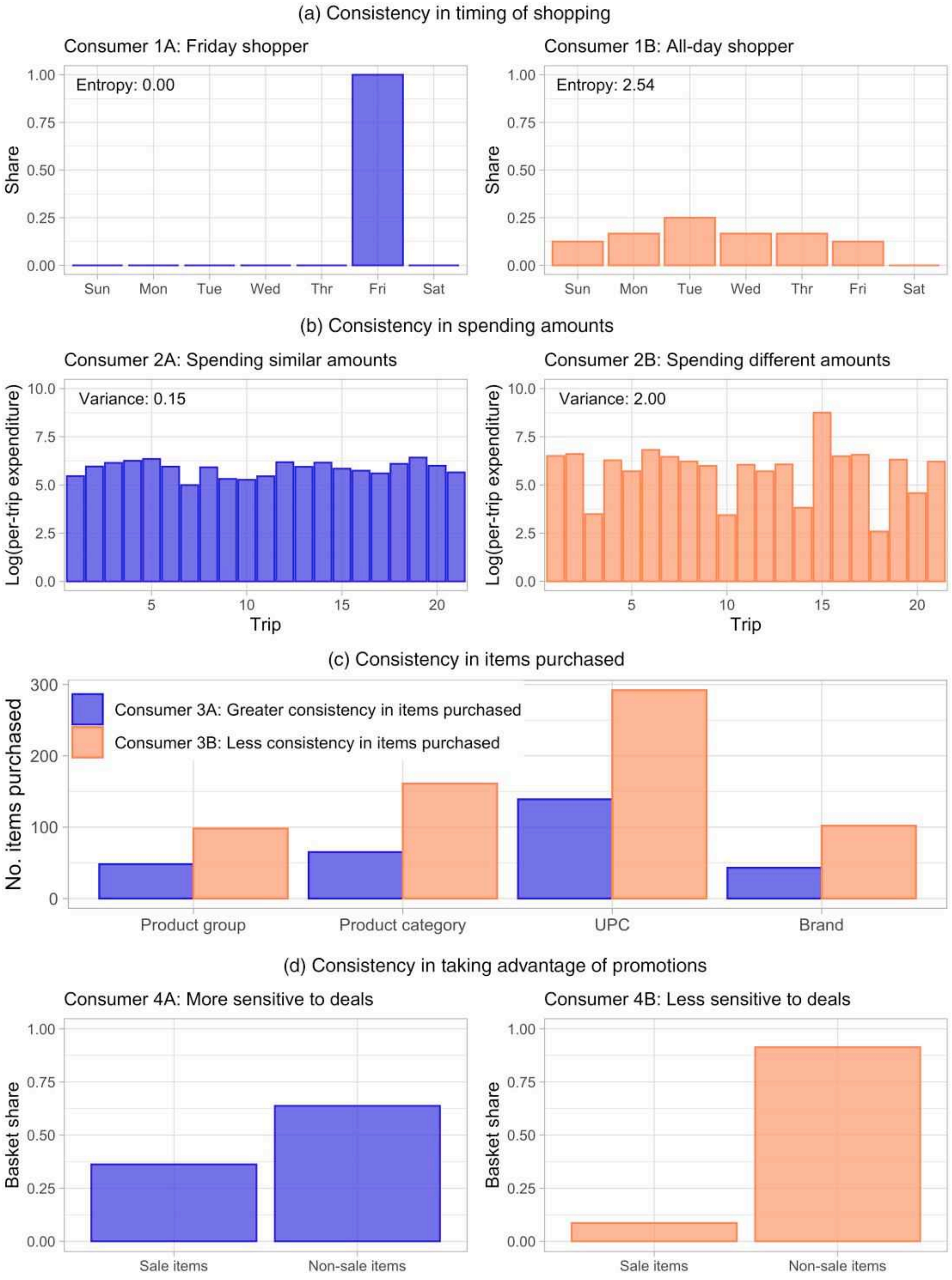
Figure 2 provides stylized examples of the four types of behaviors. Figure 2(a) shows how two sample consumers who made the same number of shopping trips to a grocery store during period 1 allocated the trips across days of the week. Consumer 1A always visited the grocery store on Fridays, whereas consumer 1B shopped on various days. Figure 2(b) compares how two sample consumers who made the same number of shopping trips allocated grocery expenditures across trips. Consumer 2A spent similar amounts of money on every visit, whereas consumer 2B displayed greater irregularity. Figure 2(c) compares two consumers who made the same number of item-level transactions (i.e., the number of unique UPC–trip pairs) during period 1. Compared with 3B, consumer 3A purchased a smaller number of unique product groups, product categories, UPCs, and brands. In other words, consumer 3A was more persistent in category and brand choices over time. Figure 2(d) compares two consumers who made the same number of item-level transactions in period 1. When comparing the items in the grocery baskets pooled across all trips, we observe that a larger share of the items in consumer 4A's baskets were on sale at the time of purchase relative to those of consumer 4B.

We briefly discuss related literature for each of the four behaviors and assess its association with credit risk.

3.2.1. Consistency in Timing of Shopping.

The consistency of when to shop varies across shoppers (Figure 2(a)). Guidotti et al. (2015) document that supermarket shoppers who are more consistent in their shopping times in terms of the day of the week and the time of

Figure 2. (Color online) Illustrative Examples of Grocery Shopping Habits



Note. Each plot shows two grocery shoppers in the final sample who display heterogeneous patterns of behavior in relation to each of the four grocery shopping habits of how to buy.

the day tend to generate more revenue in a supermarket than those who are less consistent.

We quantify the variability using the entropy measure developed by Shannon (1948). For a random variable whose possible outcomes x_i occurs with probability $P(x_i)$, entropy is defined as $\text{Entropy}(X) = -\sum_{i=1}^{n} P(x_i)\log P(x_i)$. To construct the day-of-week trip entropy for each consumer, we compute the probability of shopping trips across days of the week during period 1 (which corresponds to $P(x_i)$ in the expression). The more random or spread out the shopping trips are across the days of the week, the larger the value of entropy.

Table 4, panel A, reports the average marginal effects of the entropy measure from a regression in which the day-of-week trip entropy is used as the focal grocery variable in Equation (1). We include the log of total trip frequency in the regression as an additional control as the entropy measure is sensitive to the number of observations used to compute it. The first row of the table shows that a standard deviation increase in the trip entropy is associated with a 4.3 percentage point increase in the probability of being a sloppy payer (over a base probability of 12%) and a 4.1 percentage point increase in the probability of being a defaulter (over a base probability of 7%). On the other hand, the same change in the trip entropy is negatively associated with the probability of being a never-delinquent.[9]

We also examine whether across-segment variation in trip timing is robust to another measure: the variability of what times of day a consumer shops. To construct an entropy measure similarly, we bin times of day into six distinct four-hour blocks (2–5 a.m., 6–9 a.m., 10 a.m.–1 p.m., 2–5 p.m., 6–9 p.m., 10 p.m.–1 a.m.). A consumer's trip probabilities in the six blocks are used to compute the time-of-day entropy variable. A similar pattern emerges. The second row of Table 4, panel A, shows that a one standard deviation increase in the time-of-day entropy is associated with a 2.3 percentage point increase in the likelihood of being a sloppy payer or a defaulter.

3.2.2. Consistency in Spending Amounts.

Individuals differ in the variability of grocery expenditure over time (Figure 2(b)). One explanation proposed to explain such variation is the nature of shopping trip goals, which can be defined in terms of a shopping budget or a shopping list. Individuals who have an abstract (as opposed to concrete) shopping goal are found to be more responsive to in-store marketing stimuli (Lee and Ariely 2006) and engage in unplanned purchases (Bell et al. 2011). Empirically, having a concrete shopping goal may manifest through greater consistency in expenditure over time.

To capture the strength of consistency in spending amounts, we construct two proxies: (1) variance of log monthly grocery expenditure of a consumer and (2) variance of log per-trip grocery expenditure of a consumer.[10] Using each of these two proxies as the focal grocery variable, we run the regression in Equation (1). The log of total grocery expenditure is included as an

Table 4. Grocery Habits by Credit Card Segment

Grocery habit G_i	Never-delinquents	Sloppy payers	Defaulters
Panel A. Consistency in the timing of shopping			
Day-of-week trip entropy	−0.084***	0.043***	0.041***
	(0.008)	(0.006)	(0.005)
Time-of-day trip entropy	−0.046***	0.023***	0.023***
	(0.007)	(0.005)	(0.004)
Panel B. Consistency in spending amounts			
Variance of log monthly grocery expenditure	−0.009***	0.006***	0.004***
	(0.001)	(0.001)	(0.001)
Variance of log per-trip grocery expenditure	−0.005	0.002	0.003
	(0.004)	(0.003)	(0.002)
Panel C. Consistency in items purchased			
Log of the number of unique product groups purchased	−0.016	0.027***	−0.010
	(0.012)	(0.010)	(0.008)
Log of the number of unique product categories purchased	−0.033***	0.045***	−0.011
	(0.013)	(0.011)	(0.008)
Log of the number of unique UPCs purchased	−0.050***	0.074***	−0.024**
	(0.017)	(0.014)	(0.011)
Log of the number of unique brands purchased	−0.001	0.023**	−0.022***
	(0.011)	(0.009)	(0.007)
Panel D. Consistency in taking advantage of promotions			
Basket share of sale UPC	0.339***	−0.234***	−0.105***
	(0.025)	(0.021)	(0.016)
Expenditure share of sale UPC	0.138***	−0.089***	−0.049***
	(0.017)	(0.014)	(0.011)
Pr(buy sale UPC | buy category)	0.288***	−0.205***	−0.083***
	(0.022)	(0.019)	(0.014)

Notes. The table shows the average percent change in the probability of being a given segment for a standard deviation increase in the focal grocery habit variable. Standard errors (in parentheses) were computed using the delta method. Each row reports the estimation results of a separate regression. Results from the joint regression that includes all how-to-buy grocery features are reported in Online Appendix A.3.

Significance level: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

additional control so that individuals with similar levels of grocery spending are compared.

Table 4, panel B, reports the results from the two regressions. The first row shows the average marginal effects of the variance of log monthly grocery expenditure. A standard deviation increase in the variance reduces the probability of being a never-delinquent by 0.9 percentage points, whereas it raises the probability of being a sloppy payer and a defaulter by 0.6 and 0.4 percentage points, respectively. The variance of log per-trip grocery expenditure reveals similar patterns: spending consistent amounts on each trip is positively correlated with always paying credit card bills on time, whereas it is negatively associated with missing credit card payments or defaulting.

3.2.3. Consistency in Items Purchased. Shoppers have different varieties of items in their grocery baskets (Figure 2(c)). Researchers offer various behavioral and psychological explanations for observed persistence in choices, which include brand loyalty (Guadagni and Little 1983), addiction (Gordon and Sun 2015), habit persistence (Heckman 1981), learning (Osborne 2011), heterogeneity in intrinsic preference (Trijp et al. 1996), utility from variety itself (Ratner et al. 1999), psychological switching costs (Farrell and Klemperer 2007), product attribute satiety (McAlister 1982), and psychological stimulation and arousal (Menon and Kahn 1995).

To operationalize consistency of choices in brands and categories, we use a simple measure of the number of unique items purchased, in which an item is defined at four levels: product groups, product categories, UPCs, and brands. Table 4, panel C, reports the estimation results from the four regressions, each of which uses one of the four definitions as the focal grocery variable. The log of the number of item-level transactions is included as an additional control. Across the board, we find sloppy payers are more likely to exhibit variety-seeking behavior: they tend to have a greater breadth of product groups, product categories, UPCs, and brands in their baskets relative to the other two segments. This variety-seeking behavior does not appear to be primarily driven by deal-seeking behavior as we demonstrate next in our discussion on the consistency in taking advantage of promotions. In contrast, never-delinquents and defaulters exhibit greater persistence in their category and brand choices.

3.2.4. Consistency in Taking Advantage of Promotions. Between-individual variation in the tendency of taking advantage of deals and discounts (Figure 2(d)) has been attributed to various behavioral explanations, including inattention to price and promotions (Dickson and Sawyer 1990), price search (Walters 1991, Urbany et al. 1996), psychographic characteristics (Ailawadi et al. 2001), and concreteness of shopping goals (Bell et al. 2011).

To explore correlations between individuals' propensity to leverage deals and credit risk, we use the supermarket data on temporary price discounts from which we observe whether a given UPC was on sale at the time a consumer purchased it. As a proxy for deal sensitivity, we compute the share of total grocery transactions made on sale items. The first row of Table 4, panel D, reports the result of the regression based on the proxy (log of total grocery expenditure is included as an additional control). Never-delinquents are likely to have a larger share of sale items in their baskets relative to sloppy payers and defaulters. Specifically, a one standard deviation increase in the share of sale items is associated with a 33.9 percentage point increase in the probability of being a never-delinquent (over a base probability of 81%). Results are qualitatively similar when we look at the expenditure share rather than transaction share: increasing the expenditure share of sale items by one standard deviation increases the probability of being a never-delinquent by 13.8 percentage points.

We turn to a different measure of consistency in taking advantage of deals, which is the probability of buying a UPC that was on sale at the time of purchase, conditional on visiting a store and buying a given category. As reported in the third row of Table 4, panel D, we find that increasing the conditional probability of buying a sale item by one standard deviation increases the probability of being a never-delinquent by 28.8 percentage points.[11] Combined with the lower variety-seeking behavior of never-delinquents documented earlier, this may suggest that never-delinquents are more likely to time their purchases of preferred categories or brands to take advantage of deals. On the other hand, increasing the probability of buying sale items by a standard deviation leads to the reduction of the probability of being a sloppy payer and a defaulter by 20.5 and 8.3 percentage points, respectively.

3.3. Potential Mechanisms

We find a significant correlation between credit risk and what and how consumers buy in a grocery store even after controlling for credit scores, income, and other sociodemographic variables. Although we do not intend to test specific theories in our data, we provide suggestive evidence that may shed light on the potential mechanisms behind the observed correlations.

3.3.1. What to Buy. We explore the correlation between one's credit risk and one's overall shopping basket, which differs from our previous analysis conducted at the item level. To characterize a shopping basket, we hired workers from Amazon Mechanical Turk (MTurk) to rate the items in Table 3 along two dimensions:

healthiness and convenience (see Online Appendix A.4 for details on the survey). The healthiness of consumer i's basket is defined as the sum of the normalized item-level survey ratings (between zero and one) with various weights:

$$\text{Healthiness of Basket}_i = \sum_{k=1}^{K} \omega_{ik} \cdot \text{Normalized Healthiness Rating}_k. \quad (2)$$

The subscript k represents a food item,[12] and ω_{ik} is item k's weight, for which we use consumer i's share of expenditure on item k. The convenience of a basket is defined similarly. We then correlate the healthiness and convenience of consumer i's baskets to the segment using Equation (1).

Table 5 presents the marginal effects of the healthiness and convenience of a basket. When considering healthiness and convenience separately in the first two columns, we find that a one standard deviation increase in the healthiness score of a grocery basket is associated with a 27.5 percentage point increase in the probability of being a never-delinquent. Similarly, a one standard deviation increase in the convenience score of a basket is associated with a 13.3 percentage point increase in the probability of being a never-delinquent.

Note that the first two regressions do not control for the possibility that some items may be similar in convenience but differ in healthiness and vice versa. For instance, processed foods (typically perceived as unhealthy) and salads (typically perceived as healthy) can both be considered high-convenience items. The last column in Table 5 shows the marginal effect from a joint regression of the healthiness and convenience scores, and this allows us to measure the marginal effect of healthiness (convenience) after controlling for convenience (healthiness). Whereas the patterns for healthiness remain largely the same, the direction of the effects of the convenience score flips. Defaulters tend to allocate a greater share of their grocery expenditure to more convenient items. For sloppy payers, the evidence is weaker both statistically and in magnitude but directionally the same as defaulters. In addition, when we use purchase frequency share as a weight in Equation (2) instead of expenditure share, we find qualitatively similar patterns with even greater statistical significance.

Table 5. Correlations Between Healthiness and Convenience of Basket and Credit Risk

	(1) Healthiness only	(2) Convenience only	(3) Healthiness & convenience
Healthiness			
Never-delinquents	0.275***		0.314***
	(0.038)		(0.051)
Sloppy payers	−0.120***		−0.120***
	(0.032)		(0.042)
Defaulters	−0.155***		−0.194***
	(0.026)		(0.033)
Convenience			
Never-delinquents		0.133***	−0.052
		(0.034)	(0.044)
Sloppy payers		−0.074***	0.001
		(0.028)	(0.037)
Defaulters		−0.059***	0.052*
		(0.022)	(0.027)

Notes. Table shows the average percentage change in the probability of being a given segment for a standard deviation increase in the focal variable. Each column reports the estimation results of a separate regression. Standard errors (in parentheses) were computed using the delta method.

Significance level: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Overall, we find that, compared with sloppy payers and defaulters, never-delinquents tend to have a grocery basket that is healthier and less convenient.

3.3.2. How to Buy.

One potential explanation for the inconsistency in grocery shopping behaviors at the focal grocery store is that consumers shop around at different stores, potentially searching for better prices or deals. For example, if a consumer only shops at Walmart when the chain offers price promotions for a particular category, the consumer might appear to be inconsistent in visits to Walmart. We investigate the extent to which the consistency (or lack thereof) in grocery shopping behaviors can be explained by the tendency to shop at multiple stores.

To this end, we construct a proxy for a consumer's switching across stores. Specifically, we calculate a consumer's store choice entropy using our credit card spending data, which records transactions at various grocery stores or supermarkets. Similar to the previously mentioned day-of-week and time-of-day trip entropy, the store choice entropy is highest when the consumer makes purchases randomly across stores and lowest when the consumer always shops at the same store.

We then examine the relationship between the store choice entropy and consistency in grocery shopping behaviors in a simple regression framework:

$$\text{Consistency}_i = \sum_{d=1}^{10} \beta_d \cdot 1(\text{Store Choice Entropy Decile}_i = d) + \boldsymbol{X}_i'\gamma + \varepsilon_i, \quad (3)$$

where Consistency_i is one of our consistency measures and $\boldsymbol{X_i}$ is a vector of control variables, including the number of dependents, employment status, occupation, monthly income, credit scores, average monthly grocery expenditure, and average monthly shopping frequency. We group consumers into deciles by store choice entropy to allow for a flexible relationship between the tendency to shop around and our consistency measures. β_d are the parameters of interest.

Our findings show that the tendency to shop around at different stores can partially explain the inconsistency in grocery shopping behaviors. Consumers who are more likely to shop around, particularly those in the highest store choice entropy decile, exhibit greater inconsistency in when they visit the focal grocery store and the types of items they purchase. However, these consumers do not demonstrate inconsistency in all aspects of grocery shopping. Specifically, they spend similar amounts at the focal grocery store from month to month and take advantage of deals. Together, these patterns suggest that individuals who shop around at different stores may be price-sensitive consumers with a well-defined grocery budget. Further, we find a positive association between the tendency to shop around and paying credit card bills on time. Detailed regression results are reported in Online Appendix A.5.

3.3.3. Discussion. Although we have investigated some potential explanations for our findings, a multitude of factors can explain the observed correlations between individuals' grocery shopping habits and their credit card payment behaviors. One such factor is differences in time constraints, which may not be fully captured by the sociodemographic variables we explicitly consider. For instance, relative to individuals with multiple part-time jobs, those with stable jobs and fixed work schedules may have more time for home cooking and rely less on high-convenience items. These disparities in employment conditions may also explain their ability to pay bills on time.

Psychological factors can also offer an explanation. Buying healthy items may be attributed to one's ability to exercise self-regulation, which involves the choice of long-term benefits over short-term pleasures. Conversely, failing to make credit card payments may be indicative of low self-regulation given its association with negative long-term consequences, including reduced credit scores and unfavorable contract terms.

Moreover, these two explanations, time constraints and self-regulation, may not be unique or mutually exclusive as skills required for obtaining full-time employment may be correlated with self-regulation.

We note that it is beyond the scope of this study to test directly to what extent the observed correlations are attributed to a particular explanation. As shown in the earlier examples, they could be because of some vertical constraints (e.g., time constraints, budget constraints), horizontal heterogeneity (e.g., personality traits, cognitive abilities), other unobservables, or a combination of these. Additional data, such as data from laboratory experiments, would be necessary to investigate these different drivers of credit risk. We leave these possibilities for future studies.

4. Integrating Grocery Data into Credit Risk Prediction

This section asks the extent to which knowledge of one's grocery habits is incrementally helpful in predicting credit card payment behavior relative to traditional risk measures such as income and credit scores. To this end, we build and compare the predictive power of two credit scoring algorithms that assume different information sets of the lender: one that uses only traditional risk measures and the other that incorporates grocery data in addition to these traditional measures. We hold the modeling approach constant to isolate the incremental change in predictive accuracy resulting from the use of grocery data for a given approach.

4.1. Building Credit Scoring Algorithms

We formulate the problem as a supervised learning task in which we first learn a model of the relationship between consumer data (features) and observed payment behaviors (outcome variables) in a given training set. We maintain the two-period design in Figure 1: consumer data in period 1 are used to construct features, and credit card payment behavior in period 2 is used to construct outcome variables. Once the relationship is learned from a training set, we use the model to predict payment behaviors for consumers in a holdout set based on consumer data available at the time of prediction. We describe how consumer data are transformed into inputs for building our scoring algorithms.

4.1.1. Input Features. Three types of consumer data are considered: sociodemographic variables, credit scores, and grocery data. We preprocess sociodemographic variables by converting categorical variables (i.e., number of dependents, employment status, occupation) into a set of binary indicators to be used as input features in our credit scoring algorithms. Continuous sociodemographic variables are used directly as inputs without discretization.

Recall that some consumers in our sample do not have credit scores. We train separate credit scoring algorithms for consumers with and without credit scores to allow for the potential differential impact of grocery data on each group. For scored consumers, we include credit scores as a continuous input feature, whereas for unscored consumers, we do not use credit scores as an input feature.

To transform grocery data into usable inputs, we create a set of features that capture the five types of grocery habits examined in Section 3 using the same metrics. In addition, we create another set of features that characterize grocery shopping intensity, such as total grocery expenditure, total shopping trips, and number of item-level transactions. Summary statistics of the resulting input features for various types of consumer data are reported in Online Appendix B.1.

4.1.2. Outcome Variables and Three-Class Classification. Consistent with our consumer segmentation strategy, we incorporate the three outcome variables, or classes, into the credit scoring algorithm: never-delinquents, sloppy payers, and defaulters.

A commonly used framework for credit risk prediction assumes two classes: defaulters and nondefaulters. Within this framework, the outcome variable is an indicator of whether an applicant failed to repay the principal, and outputs from such models are interpreted as default probabilities.[13] This framework suffices to characterize problems of lenders, such as mortgage and auto loan lenders, whose expected profits depend mostly on whether a borrower defaults or not. By contrast, for the credit card industry, in which a sizeable portion of revenue comes from delinquent but not defaulting cardholders, a finer risk segmentation may have immediate profit implications. Our chosen segmentation is one of the schemes used by our data sponsor, which can allow them to improve profits by conditioning the level of service on the predicted consumer segment.[14] The three-class classification framework allows us to incorporate this institutional feature into prediction.

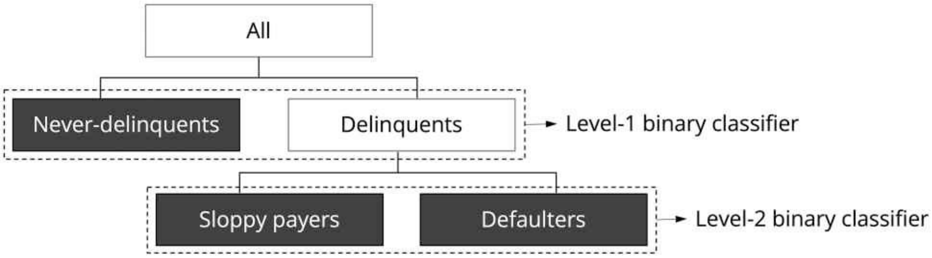
There are two general approaches to performing multiclass classification: flat and hierarchical. One of the most commonly used flat approaches is the one-versus-all approach, which involves transforming the problem into a set of binary classification tasks, each of which consists of a classifier that separates each class from all other classes. By contrast, the hierarchical approach partitions the class space into a predefined hierarchy and trains a binary classifier at each level of hierarchy. Our preferred approach is a hierarchical approach that assumes a tree-like hierarchy and decomposes the prediction problem into two binary classification problems, each of which corresponds to a split in the tree as illustrated in Figure 3.[15] The first problem is to separate never-delinquents from delinquents, and delinquents are a meta-class that includes both sloppy payer and defaulter classes. The second problem is to discriminate, among delinquents, between sloppy payers and defaulters.[16]

4.1.3. Learning Algorithms. To fit a classifier for each level of the problem, we use XGBoost (eXtreme Gradient Boosting; Chen and Guestrin 2016). XGBoost is a nonlinear and nonparametric estimator that builds and combines a set of tree-based weak learners, which are characterized by high bias and low variance, into a strong learner with low bias and low variance via boosting. Note that, as our empirical strategy is to hold the classifier constant and vary information sets, any learning algorithm other than XGBoost could be used.

4.1.4. Model Evaluation. We use a nested cross-validation approach to quantify the predictive power of credit scoring algorithms, which involves two cross-validation loops (Figure 4). In the outer loop, the sample consumers are first partitioned into 10 roughly equal-sized, nonoverlapping folds.[17] For each test fold k, the remaining nine folds serve as a training set based on which the relationship between the features and the outcome variable is learned and the corresponding hyperparameters are tuned. To find the optimal model parameters, in the inner loop, the training set is partitioned again into 10 folds based on which the usual 10-fold cross-validation is performed. The out-of-sample predictive accuracy of the resulting model is tested on the held-out test fold k. We repeat this procedure by running through each test fold to generate 10 predictive accuracy measures. To summarize, we leverage independent subsets of data for model selection (inner loop) and model assessment (outer loop), reducing the chance of overfitting or biasing model assessment (Cawley and Talbot 2010).

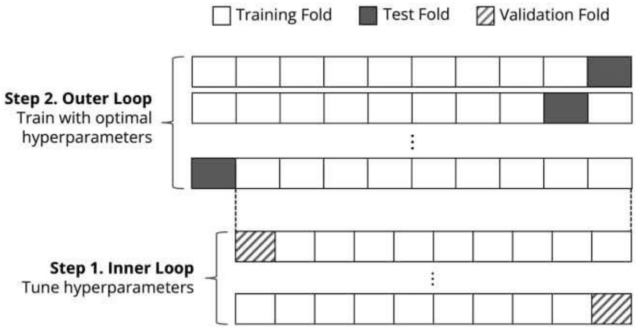
We report the distribution of the predictive accuracy measures computed on the 10 test folds. This way, we can examine the extent to which estimated predictive power is robust across partitions of the data. As our main evaluation metric, we use the AUC for two main

Figure 3. Hierarchical Structure of Consumer Segments



Notes. We decompose our prediction problem into two binary classification problems, each corresponding to a split in the hierarchy of consumer segments. The first problem is to separate never-delinquents from delinquents, and delinquents are a meta-class that includes both sloppy payer and defaulter classes. The second problem is to discriminate, among delinquents, between sloppy payers and defaulters.

Figure 4. Nested Cross-Validation for Model Learning



reasons. First, this metric allows us to compare the predictive power of algorithms without requiring us to take a stance on a particular discrimination threshold, which may ultimately depend on the lender's objective. Second, the AUC is a more robust metric for evaluating the predictive performance of imbalanced data. As the class distribution in our data is highly skewed with a low occurrence of missing a payment in the sample, metrics such as accuracy, precision, or recall may not accurately reflect the performance of algorithms.

4.2. Predictive Power of Grocery Data

4.2.1. Relative Predictive Power of Grocery Data. We first ask how the predictive power of grocery data compares with other data sources available to credit card issuers. Table 6 reports the out-of-sample AUCs of various data sources for scored and unscored consumers.

For consumers whose credit scores are not available, grocery data consistently and substantially outperforms sociodemographic variables, including income. When it comes to separating never-delinquents and delinquents, the out-of-sample AUC of grocery data is 0.639, which is 6.9 percentage points higher than that achieved by sociodemographic variables. In distinguishing between sloppy payers and defaulters, grocery data outperforms sociodemographic variables by 3.8 percentage points.

Table 6. Out-of-Sample AUC of Various Data Sources

	Unscored consumers		Scored consumers	
Information set	Level 1	Level 2	Level 1	Level 2
Sociodemographics	0.570	0.560	0.568	0.623
	(0.027)	(0.031)	(0.020)	(0.029)
Credit scores	–	–	0.639	0.568
	–	–	(0.027)	(0.053)
Grocery data	0.639	0.598	0.619	0.576
	(0.020)	(0.032)	(0.015)	(0.049)

Notes. The table reports the mean and standard deviation (in parentheses) of 10 AUCs based on the 10 folds for each level. Level 1 refers to a binary classifier that classifies consumers into never-delinquents and ever-delinquents. Level 2 refers to a binary classifier that classifies delinquent consumers into sloppy payers and defaulters.

For scored consumers, grocery data exhibits comparable predictive accuracy to credit scores and, in some cases, even outperforms them. Specifically, grocery data achieves 97% and 101% of the predictive accuracy of credit scores for levels 1 and 2 problems, respectively. To put this into perspective, our estimates of the predictive power of credit scores lie slightly above the estimates reported in the literature, particularly in developing markets. For instance, Björkegren and Grissen (2020) study a South American country where a telecom extends credit to mobile phone customers in the form of postpaid plans. They find that, in a sample of consumers with credit scores constructed based on thin credit files, the out-of-sample AUC of the algorithm that leverages credit bureau information is between 0.510 and 0.565.

4.2.2. Incremental Predictive Power of Grocery Data. We turn to evaluate the marginal impact of grocery data depending on the baseline information set of lenders. The availability of consumer data for lenders varies widely across markets. In developing and emerging markets, lenders often rely only on self-reported sociodemographic variables as a significant portion of consumers do not have credit scores. On the other hand, in developed markets, credit scores tend to be available for a larger fraction of consumers although certain consumer segments, such as recent college graduates, recent immigrants, and freelancers, may still lack credit scores.

To investigate the marginal value of grocery data across various markets, we examine its incremental predictive power in two scenarios. Specifically, we consider a baseline information set of lenders that includes only sociodemographic variables and another that includes both sociodemographic variables and credit scores. We then compare the incremental predictive accuracy of grocery data relative to these two baseline information sets. The results are reported in Table 7.

When consumers do not have credit scores, incorporating grocery data leads to a substantial improvement in predictive accuracy, ranging from 3.11 to 7.66 percentage points (5.69%–13.6%) in the out-of-sample AUC as shown in panel A of Table 7. When we dive deeper into the importance of individual features, we find that the predictive power of monthly income, a sociodemographic variable typically highly predictive of credit risk for scored consumers, is limited for unscored consumers. Instead, features related to grocery habits, such as share of sale UPCs in the shopping basket, whether they purchased canned meat, and the number of unique brands purchased, are more predictive of whether one misses a credit card bill. Further details on the importance of individual features are in

Table 7. Incremental Predictive Power of Grocery Data

Panel A. Unscored consumers				
	Out-of-sample AUC			
Baseline information set	Without grocery	With grocery	ΔAUC (pp.)	ΔAUC (%)
	Level 1: Never vs. ever-delinquents			
Sociodemographics	0.570	0.647	+7.66 pp.	+13.6%
	(0.027)	(0.017)	(2.59)	(5.01)
	Level 2: Sloppy payers vs. defaulters			
Sociodemographics	0.560	0.591	+3.11 pp.	+5.69%
	(0.031)	(0.048)	(4.50)	(8.26)
Panel B. Scored consumers				
	Out-of-sample AUC			
Baseline information set	Without grocery	With grocery	ΔAUC (pp.)	ΔAUC (%)
	Level 1: Never vs. ever-delinquents			
Sociodemographics	0.568	0.629	+6.07 pp.	+10.7%
	(0.020)	(0.015)	(1.34)	(2.61)
Sociodemographics + credit scores	0.654	0.679	+2.51 pp.	+3.91%
	(0.033)	(0.025)	(1.48)	(2.37)
	Level 2: Sloppy payers vs. defaulters			
Sociodemographics	0.623	0.630	+0.665 pp.	+1.08%
	(0.029)	(0.038)	(2.55)	(4.02)
Sociodemographics + credit scores	0.637	0.641	+0.359 pp.	+0.614%
	(0.036)	(0.043)	(2.99)	(4.81)

Notes. The table reports the mean and standard deviation (in parentheses) of 10 AUCs based on the 10 folds for each level. Level 1 refers to a binary classifier that classifies consumers into never-delinquents and ever-delinquents. Level 2 refers to a binary classifier that classifies delinquent consumers into sloppy payers and defaulters. pp. indicates percentage points.

Online Appendix B.2. These findings highlight the potential value of utilizing grocery data in assessing credit risk for consumers who lack credit scores.

Incorporating grocery data continues to improve the predictive accuracy for consumers who have credit scores as shown in panel B of Table 7. The improvement is particularly notable in distinguishing between never-delinquents and delinquents, increasing the out-of-sample AUC by 2.51 to 6.07 percentage points (3.91% to 10.7%). The improvement is positive yet more modest in separating sloppy payers and defaulters, ranging from 0.359 to 0.665 percentage points (0.614% to 1.08%). To provide a rough sense of magnitude, we compare the improvement for the level 1 problem to the estimates found in an arguably more developed market. In the context of a Berlin-based e-commerce platform that allows users to purchase furniture on credit, Berg et al. (2020) examines the marginal impact of utilizing digital footprints, another type of alternative data, for credit scoring. They find a 5.3 percentage point increase in the out-of-sample AUC when digital footprints are added to the baseline information set, which consists of credit scores but not sociodemographic variables.

Overall, our findings indicate that consumers who are observationally equivalent in terms of traditional data, such as income and credit scores, but differ in grocery shopping behaviors exhibit varying levels of credit risk.[18]

4.2.3. Boundary Condition for the Value of Grocery Data.

Once credit card issuers and banks establish a lending relationship with a consumer by, for instance, issuing a credit card, they gain access to comprehensive and detailed first party data on the consumer's credit card spending patterns and repayment behaviors, which may provide more relevant information on credit risk compared with alternative data sources, such as grocery data. Lenders evaluating existing cardholders' applications for their different financial products could utilize such data to predict their credit risk. To explore the value of grocery data in this scenario, we create a new baseline information set that includes features constructed based on our credit card spending data and credit card payment history from period 1 (in addition to sociodemographic variables and credit scores) and examine the predictive gain from grocery data.

We find that the marginal effect of grocery data diminishes significantly when lenders have access to detailed credit card data. As credit card features alone are highly predictive of credit risk, further incorporating grocery data into the algorithm leads to a positive but marginal improvement in predictive accuracy for scored consumers (less than one percentage point increase for both levels 1 and 2 problems). The predictive gain is more pronounced for consumers without credit scores, which suggests a potential overlap in the signals provided by credit scores and those derived from detailed credit card data.

These findings highlight a boundary condition for using grocery data as alternative data, in which its value lies more in predicting the credit risk of new consumers, whether they are new to credit or new to the specific lender, rather than existing ones. They also provide unique empirical evidence on the value of alternative data relative to relationship-specific proprietary data owned by banks (Berg et al. 2022). More details on the analysis are provided in Online Appendix B.3.

4.3. Selection of the Approved and Enrolled Cardholders

A natural question is to what extent our findings can be generalized to a broader consumer population beyond our sample. Given that grocery data is found to be incrementally predictive of credit risk in a more homogeneous subset of the population in terms of both their ex ante credit risk and ownership of a grocery loyalty card, we speculate its predictive power would be present in the broader population. The key challenge in addressing this question more rigorously is the missing data problem: we only observe credit risk or default outcomes for credit card applicants who were approved and enrolled but not for those who were denied credit, approved but not enrolled, or did not apply for credit at all. As the probability of missing data is likely correlated with credit risk (that is, data are missing not at random because of the data provider's screening process), our sample may not be representative of the entire population.[19]

The ideal experiment for assessing the value of grocery data in the broader population would involve randomly (or extremely permissively) extending credit to both applicants and nonapplicants and tracking their repayment behaviors over an extended period of time. However, this experiment would be either prohibitively expensive[20] or impossible because of legal restrictions.[21] To tackle this challenge, alternative approaches have been proposed to infer data on the rejected, which require assumptions with varying plausibility.[22]

Given the nature of our data, we interpret our results only for the sample of applicants and do not aim to assess the value of grocery data in the general population. However, we seek to provide suggestive evidence of its predictive power in the neighborhood of the issuer's threshold risk. The intuition is as follows. We assume the current customer portfolio of the issuer is generated by the issuer's policy, in which only applicants with a predicted risk below a specific threshold are approved. Given this assumption, we can implement a hypothetical stricter approval policy by ranking the sample consumers based on their predicted risk and approving only those who fall below a given percentile of the risk distribution (i.e., approve a less risky subset of the existing cardholders). By repeating this exercise with progressively stricter policies and estimating the predictive power of grocery data in each subset, we can trace out the trajectory of its predictive power in the progressively narrower and more homogeneous population.

This exercise shows that, as the lender's policy becomes more stringent, the absolute predictive power of grocery data remains relatively stable, whereas that of other data sources tends to shrink, which makes grocery data continue to be incrementally helpful in predicting credit risk across different threshold risk levels. If the relationship between grocery data and credit risk remains consistent in the regions not observed in the data (e.g., rejected applicants), grocery data could be helpful in predicting credit risk beyond the specific population in our sample. We provide further details of the simulation in Online Appendix B.4.

5. Integrating Grocery Data into Credit Decisions

Although the use of grocery data can improve the accuracy of credit risk prediction, its impact on credit decisions and outcomes largely depends on whether and how lenders choose to utilize the predictions. This section explores the impact of leveraging grocery data in credit decisions by simulating a credit card issuer's credit extension process. In this simulation, approval decisions are made based on the credit risk predictions described in Section 4.[23]

We reiterate that our data set is limited to applicants who have applied, been approved, and enrolled in credit cards provided by the data provider. Therefore, we do not attempt to implement a decision rule that requires data on rejected applicants. Instead, we focus on a decision rule that is applicable within the population of approved applicants as we elaborate shortly.

Further, our simulation aims to illustrate the short-run impact of incorporating grocery data rather than providing a precise quantification of the welfare effect. To this end, we formulate the lender's decision as a pure prediction problem, in which the decision to approve or reject an applicant does not causally affect the default probability, either ex ante or ex post, of the applicant.[24] We designed these simulations in collaboration with our data sponsor, who confirmed that this stylized approach effectively captures the most important aspects of their credit card approval decision.

5.1. A Simple Model of Issuer's Credit Extension Decisions

Consider a credit card issuer who chooses an action $a_i \in \{0,1\}$ for credit card applicant i, where $a_i = 1$ indicates approval and $a_i = 0$ rejection. The issuer gets a payoff that depends on both the action a_i and the true state of the world ω_i, where the state corresponds to applicant i's segment. There are three states of the world, $\omega_i \in \{n, s, d\}$, where n, s, and d are a never-delinquent, sloppy payer, and defaulter, respectively. Because the state is realized after the issuer chooses an action, the issuer faces uncertainty about the state at the time of decision making. Before making the decision, the issuer produces a set of predictions, or posterior probabilities of each state, $p_i = (p_i^n, p_i^s, p_i^d)$.

Given p_i, the issuer chooses the optimal action according to its decision rule. Suppose the issuer's policy is to choose an action that yields the highest expected payoffs:

$$a_i^* = \underset{a_i \in \{0,1\}}{\arg\max}\, (p_i^n \pi_i^n + p_i^s \pi_i^s + p_i^d \pi_i^d - \delta) \cdot a_i, \tag{4}$$

where π_i^n, π_i^s, and π_i^d are the payoffs when the issuer extends credit to i and the true state is n, s, and d, respectively. We assume zero payoff of not extending credit regardless of state (i.e., zero opportunity cost of false negatives).[25] δ is the issuer's threshold payoff, representing how strict or lenient the approval policy is. A higher δ indicates a stricter decision rule.

We simulate the issuer's decisions on a_i's under different data environments. This simulation involves calibrating the model by obtaining segment probabilities p_i's and segment-specific payoffs π_i's in Equation (4). Because the outputs from the credit scoring algorithms in Section 4 cannot be directly interpreted as probabilities, we transform them into probabilities to obtain p_i's. We also calibrate the π_i's, representing segment-specific cash flows, by utilizing our credit card spending data and payment history. The data sponsor validated both the model structure and parameter values used in the simulations to ensure that the model accurately represents the key features of their decision-making process. We discuss these processes in detail in Online Appendix C.

Recall that the decision variable a_i has no direct effect on p_i's or π_i's, that is, $p_i \neq f(a_i)$ and $\pi_i \neq f(a_i)$. Thus, the lender's problem is to predict whether the default probability of an applicant, weighted by the segment-specific payoff, is high enough to reject the applicant given the threshold δ.[26]

5.2. Impact of Incorporating Grocery Data on Credit Approval Decisions

We simulate one potential approach to incorporating grocery data into credit extension decisions, which is to use it as an additional screening device. To illustrate this, consider a lender with a two-stage, sequential screening mechanism. In the initial stage, the lender evaluates applicants using only standard data (without grocery data) to approve a subset of consumers tentatively. In the subsequent stage, both standard data and grocery data are used to further weed out applicants among those who passed the first stage.

To simulate this decision rule, we assume that our sample consumers represent the set of applicants who made it through the first stage. We are agnostic about the credit scoring algorithm the lender might have used in the first stage and treat the sample as given. To simulate the second stage decision, we build a credit scoring algorithm that incorporates both standard data and grocery data and use it to evaluate which tentatively approved consumers should be rejected. Note that, in making the first stage decision, the lender must have established the minimum expected payoff required for an applicant to be approved or the threshold value δ. To back out the implied threshold in our data, we find the value that results in barely approving every sample consumer, thereby rationalizing the observed approval decisions. We then use this inferred threshold to simulate the second stage decision.

We first assess the impact of grocery data on the issuer's profitability, which can shape the issuer's incentive to utilize grocery data in its credit decisions. Table 8 presents the changes in per-person payoffs brought about by the inclusion of grocery data, in which the per-person payoff is calculated by dividing the sum of expected cash flows from the approved applicants by the number of approved applicants. In addition, the table reports the changes in the approval probability of the three customer segments with N indicating never-delinquents, S sloppy payers, and D defaulters.

For unscored consumers, incorporating grocery data into credit extension decisions leads to a 1.46% increase in per-person profit driven by the lower average credit risk of approved applicants. In particular, defaulters experience a 7.16 percentage point decrease in approval probability. Never-delinquents and sloppy payers also face a decrease in approval probability, but the reduction's magnitude is less pronounced compared with defaulters. When credit scores are available, the impact of grocery data on profitability is marginal, resulting in a 0.025% increase in per-person payoffs. This is the case despite grocery data's positive incremental predictive power for scored consumers as demonstrated in Table 7. Taken together, these findings suggest that lenders may have a stronger incentive to acquire, collect, and utilize grocery data for scoring and screening applicants who do not have a traditional credit score.

5.2.1. Discussion.

In this section, we consider a particular use case of grocery data in credit approval decisions, in which the issuer leverages grocery data as an

Table 8. Incremental Profit Introduced by Grocery Data

Panel A. Unscored consumers

Baseline information set	ΔPer-person payoff	ΔPr(Approve) *N*	ΔPr(Approve) *S*	ΔPr(Approve) *D*
Sociodemographics	1.46% (1.32)	−2.06 pp. (1.52)	−4.44 pp. (3.62)	−7.16 pp. (5.58)

Panel B. Scored consumers

Baseline information set	ΔPer-person payoff	ΔPr(Approve) *N*	ΔPr(Approve) *S*	ΔPr(Approve) *D*
Sociodemographics	1.48% (0.660)	−2.39 pp. (0.841)	−5.26 pp. (3.10)	−9.69 pp. (3.52)
Sociodemographics + credit scores	0.025% (0.079)	−0.017 pp. (0.036)	−0.142 pp. (0.300)	−0.133 pp. (0.422)

Notes. The table reports the mean and standard deviation (in parentheses) of 10 simulation results, each based on the 10 folds. The last three columns present the changes in the approval probability of three customer segments with *N* indicating never-delinquents, *S* sloppy payers, and *D* defaulters.

additional screening mechanism. We find that the use of grocery data can increase per-person profits by rejecting defaulters to a greater degree than nondefaulters.

We reiterate that our approach does not account for applicants who failed to meet the traditional lending criteria and, therefore, were excluded from our sample. However, in practice, it is at lenders' discretion to determine how they incorporate grocery data as additional input into their decision-making process, depending on their specific objectives. For example, some lenders may choose to evaluate applicants using both standard and grocery data in a single step rather than employing the two-step processes we describe. Alternatively, other lenders might use grocery data to reevaluate applicants who do not meet the traditional lending criteria but have some good grocery shopping habits, providing them with a second opportunity for credit approval. These alternative approaches allow lenders to leverage the signals of credit risk in grocery data not only for further screening (as demonstrated in our simulations), but also for approving applicants who would otherwise be denied credit in the absence of grocery data. Investigating the impact of these decision rules inevitably requires data on applicants rejected under different lenders' legacy decision algorithms, which is currently unavailable to us. These considerations call for additional research to understand how lenders' decision rules concerning the use of alternative data in credit scoring affect the distribution of credit across different consumer groups.

6. Additional Use Case of Grocery Data: Managing Existing Cardholders

Whereas our primary focus has been on the value of grocery data for customer acquisition, we also explore its additional use case in managing existing cardholders. Once selected applicants become part of their risk portfolio, credit card companies closely monitor these customers and regularly assess their credit risk using a variety of behavioral scoring algorithms. These resulting scores enable firms to implement targeted interventions promptly, including offering targeted products and promotional offers (e.g., preapproved loans) and adjusting credit terms (e.g., modifying credit limits or interest rates).

Specifically, we build an algorithm that predicts whether a cardholder will fail to make a payment again in the following month after missing a credit card payment for the first time in a given month. Recall that, in our empirical context, a defaulter is defined as a cardholder who misses two consecutive payments, whereas a sloppy payer is a cardholder who misses a credit card payment but pays it back the next month. Accordingly, this algorithm aims to distinguish between these two customer segments upon observing a one-month delinquent cardholder. This early identification of cardholder types allows credit card companies to implement more suitable proactive measures to reduce potential future losses. For example, companies can freeze the cards of those with a sufficiently high predicted likelihood of missing a payment again, sending payment reminders to those who are likely to be sloppy payers.

We posit that changes in grocery shopping habits may contain temporary credit risk signals. For example, individuals who have recently lost their jobs might shift spending from fresh vegetables to frozen vegetables because of financial hardships. Meanwhile, among one-month delinquent cardholders, some may have missed their bills because of travel or oversight, resulting in the absence of behavioral changes or different types of changes. Considering that consumers frequently make choices in the grocery retail domain, using grocery data may enable lenders to capture these adjustments over time, which can serve as signals of temporary credit risk.

To capture these shifts in shopping habits, we create two sets of features based on grocery data. The first set aims to capture a cardholder's grocery shopping habits during regular times, referred to as permanent behaviors. These features are derived from grocery data spanning from two to six months prior to the month of the cardholder's first delinquency, following the approach outlined in Section 3. The second set of features illustrates the cardholder's shopping behaviors closer to the first delinquency month, referred to as recent behaviors. These features are constructed using grocery data from the month of the first delinquency and the preceding month. By incorporating both sets of features into the algorithm, we aim to leverage both permanent shopping habits and any deviations in these habits leading up to the delinquency event for prediction.

In this customer management problem, we utilize a different subset of our data compared with that employed in the customer acquisition problem. Specifically, we focus on cardholders who missed at least one credit card payment during the sample period and experienced their first delinquency between July 2017 and January 2019. This approach allows us to maximize the sample size used for algorithm training, ensuring that a minimum of seven months' worth of grocery data are available for prediction purposes. Our final sample consists of 7,449 cardholders, among whom 1,520 defaulted or missed two consecutive payments (defaulters), whereas 5,929 paid their bills the following month (sloppy payers). The algorithm is trained and tested using XGBoost, employing a nested cross-validation approach as illustrated in Section 4. Additional details about the data and resulting features are in Online Appendix D.

Table 9 presents the ability of individual data sources to distinguish between defaulters and sloppy payers when leveraged in isolation, measured in terms of out-of-sample AUC. The table includes credit card data as one of the data sources. The key difference between the current customer management problem and the earlier acquisition problem lies in the data sources available to the issuer at the time of prediction. When evaluating new applicants, issuers typically have access to only self-reported demographic information and credit bureau data. However, once an applicant is approved and establishes a relationship with the issuer, the issuer gains access to a wealth of internally generated data, including detailed credit card spending data. Accordingly, we apply the same feature engineering approach to credit card spending data as grocery data, focusing on creating features that reflect permanent and recent behaviors.

Table 9. Out-of-Sample AUC of Different Data Sources

Information set	Unscored consumers	Scored consumers
Sociodemographics	0.596 (0.043)	0.620 (0.025)
Credit scores	–	0.535 (0.047)
Credit card data		
Permanent behaviors	0.584 (0.028)	0.597 (0.049)
Recent behaviors	0.638 (0.032)	0.688 (0.025)
Permanent + recent behaviors	0.640 (0.026)	0.712 (0.037)
Grocery data		
Permanent behaviors	0.574 (0.047)	0.572 (0.027)
Recent behaviors	0.604 (0.029)	0.644 (0.030)
Permanent + recent behaviors	0.657 (0.015)	0.652 (0.037)

Note. The table reports the mean and standard deviation (in parentheses) of 10 AUCs based on the 10 folds.

We find that the predictive accuracy of sociodemographic variables and credit scores is limited compared with that of credit card data and grocery data, which may be attributed to their static nature. Among the features derived from credit card data, recent features consistently outperform permanent ones with a substantial performance gap, ranging from 5.4 to 9.1 percentage points. Incorporating both sets of features leads to a relatively small yet positive improvement in predictive accuracy. A similar pattern emerges with grocery data: recent grocery features outperform permanent grocery features by 3.0 to 7.2 percentage points. These findings highlight the significant advantage of observing more recent behaviors, whether in credit card spending or grocery shopping, in the context of managing delinquent accounts.

Table 10 reports the incremental predictive power of grocery data with different assumed baseline information sets. Our results show that grocery data can be valuable in predicting default among delinquent cardholders when the only information available to the issuer is the sociodemographic variables and/or credit scores of cardholders. This scenario reflects the limited information available to issuers when dealing with newly acquired credit card customers who have missed their payment in their initial month. Specifically, the inclusion of grocery data improves the out-of-sample AUC by 3.11 percentage points for unscored consumers and 7.78 percentage points for scored consumers (assuming the baseline information set comprises sociodemographic variables and credit scores).

However, the incremental value of incorporating grocery data is marginal once the issuer can access detailed credit card spending data. For both unscored or scored consumers, we find that the incremental predictive accuracy introduced by grocery data falls sharply to 0.333 and 1.18 percentage points, respectively, once features

Table 10. Incremental Predictive Power of Grocery Data

Panel A. Unscored consumers

	Out-of-sample AUC			
Baseline information set	Without grocery	With grocery	ΔAUC (pp.)	ΔAUC (%)
Sociodemographics	0.596	0.627	+3.11 pp.	+5.33%
	(0.043)	(0.041)	(2.16)	(3.92)
Sociodemographics + credit card data	0.657	0.660	+0.333 pp.	+0.507%
	(0.015)	(0.028)	(2.41)	(3.71)

Panel B. Scored consumers

	Out-of-sample AUC		ΔAUC (pp.)	ΔAUC (%)
Baseline information set	Without grocery	With grocery		
Sociodemographics	0.620	0.692	+7.20 pp.	+11.7%
	(0.025)	(0.028)	(2.82)	(4.87)
Sociodemographics + credit scores	0.622	0.700	+7.78 pp.	+12.5%
	(0.029)	(0.041)	(2.68)	(4.50)
Sociodemographics + credit scores + credit card data	0.726	0.738	+1.18 pp.	+1.65%
	(0.034)	(0.041)	(2.91)	(4.08)

Notes. The table reports the mean and standard deviation (in parentheses) of 10 AUCs based on the 10 folds. pp. indicates percentage points.

derived from credit card spending data are included in the information set. This suggests that signals derived from grocery data largely overlap with those from credit card spending data.

7. Managerial Implications

Our findings have direct managerial implications for lenders as using grocery data for credit scoring presents an opportunity to access a vast, untapped market. Lenders can expand their customer base and improve their profitability by extending credit to consumers who are currently unserved or underserved by the traditional credit system. In addition, this study introduces a new avenue for data monetization for retailers. Our data sponsor stated, “Seeing the model results opened our eyes to the possibility of using grocery data in lending.” In particular, the possibility of serving consumers who lack credit using this approach is an intriguing business case that our data sponsor will continue to explore. Further, lenders’ use of grocery data can contribute to enhancing consumer welfare by allowing them to make longer term investments in key areas, including education (Solis 2017), housing (Barakova et al. 2003, Acolin et al. 2016), and career development (Evans and Jovanovic 1989, Del Boca and Lusardi 2003).

One may wonder what makes grocery data attractive relative to other potential alternative data sources. We argue that grocery data offers several advantages when it comes to capturing individuals’ purchasing (and consumption) habits, which can be a strong predictor of their financial habits. First, as groceries are nondurable necessities, a large fraction of the consumer population shops in a grocery store and makes repeated and frequent choices in the domain. This feature enables a more reliable inference of individuals’ habits for a substantial fraction of consumers. Second, individuals make choices in a broad cross-section of product categories when grocery shopping, which allows us to observe their choices in various contexts. This aspect can help identify several habits that contain relevant signals pertaining to credit card payment behaviors. Third, grocery data capture individuals’ actual choices rather than relying on their stated intentions or self-reported behaviors. By reflecting on the trade-offs individuals face when making purchase decisions, grocery data may provide a more precise understanding of their preferences than other data sources that rely on hypothetical scenarios.

We recognize that these benefits may not be unique to grocery data. For instance, gasoline purchases may capture some meaningful aspect of individuals’ habits (e.g., gasoline trip frequency, amount of gasoline purchased, octane choice, store choice) as consumers make repeated and frequent transactions in the domain. One comparative disadvantage of these data, however, would be that they only involve a single product category (i.e., gasoline) unlike typical grocery data, which encompasses choices across multiple product categories. Another type of data that could be useful is the data on cellphone signals. We argue that, although firms may learn about the locomotion of consumers on a second-to-second basis, it would be more difficult to directly observe purchase or consumption from the data.

The presence of incremental value of across-domain data also speaks to questions concerning the boundaries of the firm. Since the 1950s, when many corporations diversified their portfolios through mergers and acquisitions, there has been a long debate over the efficiency of a conglomerate in operating unrelated businesses relative to those segments operating as stand-alone units. Some economists may find it hard to rationalize why a credit card company and a grocery store would be owned by the same firm when the two appear largely unrelated to each other at first glance (e.g., Myerson 1982, Jensen 1986, Berger and Ofek 1995). Our results provide empirical evidence for strong informational synergies between the two entities that provide a rationale for this ownership structure. This type of gain is of particular relevance in today's business environment, in which vertical and horizontal integration over seemingly distant domains is becoming increasingly representative of the industry movement.

Another relevant question is whether the ownership of both a bank and a retailer is a necessary condition for using first party retail data in lending decisions as observed in this study's empirical setting. Given the potential synergies between consumer data from different domains, it would not be surprising if the market structure were to evolve around these incentives in the long run. In fact, there has been considerable speculation that Apple will aim to become a "full-stack bank" with its financial infrastructure for payments and lending (Shevlin 2019). Apple has been gradually building its own financial infrastructure, launching its mobile payment service (Apple Pay) in 2014; offering credit cards (Apple Card) in 2019; introducing a savings account service in 2023 in partnership with Goldman Sachs; pairing up with Affirm in 2021 to offer buy now, pay later services (Apple Pay Later); and acquiring a fintech startup, Credit Kudos, that provides credit scores using open banking technology in 2022. The tech giant may leverage its detailed phone usage and location data collected through smartphones to make informed lending decisions. Similarly, whereas Amazon currently issues credit cards in partnership with Chase Bank and does not assume the risk as a lender, it has the potential to become a bank that leverages its first party transaction data from Whole Foods Market and other domains it operates in for credit scoring. In summary, firms with exclusive access to their own consumer data may enjoy a competitive advantage in the credit market.

It may appear to be more challenging to collect detailed grocery data in developing and emerging markets, in which consumers mostly shop at mom-and-pop grocery stores using cash rather than at large-scale supermarkets that systematically collect point-of-sale data. However, in fact, the retail sector is quickly becoming modernized and digitized in many such markets. For instance, mobile payment service and digital wallet adoption rates by both consumers and retailers are high in Kenya and Ghana (Collins 2019). This trend may allow and even incentivize retailers to invest in more systematic consumer data collection, especially if they could monetize their own consumer data by sharing or selling them to third party firms.

8. Conclusions

This study shows that consistent shopping behaviors exhibited in a grocery store are predictive of credit card payment behaviors above and beyond standard data sets used by lenders. Our simulations further demonstrate that such incremental predictive power can benefit both lenders and traditionally underserved consumers in credit markets, such as consumers who lack credit scores. This finding further suggests that grocery data may advance the financial inclusion of consumers with limited or no access to credit.

We caution that our findings are based on a selected sample of approved and enrolled credit card applicants. The data limitation precludes us from assessing the impact of grocery data on predictions and decisions on the broader consumer population. Further, our estimates of the impact are rather illustrative, short-run effects of using grocery data when behaviors of competing firms, consumers, and regulators are held fixed. Indeed, several theoretical papers explore the implications of strategic consumers in the context of data sharing, focusing on their incentives to reverse engineer and game the system or to withhold their data altogether in expectation of unfavorable decisions (e.g., Ball 2019, Hu et al. 2019, Frankel and Kartik 2022) with some proposing an estimator that is robust to such manipulation (e.g., Björkegren et al. 2020). Overcoming these limitations may require a new data source, ideally from an experiment. We view our work as a proof of concept that can provide insights into the design and implementation of future studies.

We leave for future research the question of how the use of grocery data or, more broadly, any first party data from nonfinancial domains, should be regulated in consumer lending markets. Relevant questions to explore include, but are not limited to the following: How can we ensure the fair and equitable distribution of lending outcomes across different consumer groups? Is it ethically justifiable to deny credit based on an applicant's purchase of cigarettes? How can we strike a balance between the utility from credit access and the disutility from privacy concerns? Given the far-reaching implications of credit access on households' life cycles and across generations, it is critical to understand the normative implications of our findings to prevent unintended negative consequences.

Acknowledgments

The authors are grateful to Brett Gordon, Brian Sternthal, Anna Tuchman, and Caio Waisman for their insightful comments. The authors have greatly benefited from the feedback of participants at the 2022 Boulder Summer Conference on Consumer Financial Decision Making, the 2022

AIM Conference, the 2022 AMA Summer Academic Conference, the 2022 TCR-AMA Impact Festival, and the 2022 BSPA Annual Conference. The authors also thank the seminar participants at the Georgia Institute of Technology, Massachusetts Institute of Technology, National University of Singapore, Northwestern University, Rice University, and the University of Notre Dame. The authors appreciate the support of Wharton Customer Analytics and an anonymous data provider for sharing their data and domain knowledge. This paper constitutes part of Lee's doctoral dissertation.

Endnotes

[1] Habits are extensively studied in psychology, economics, and marketing and are commonly defined as repeated behaviors within specific contexts. Despite the similarity in the behavioral definition of habits, scholars from various disciplines offer distinct perspectives on their interpretation. Psychologists commonly characterize habits as automatic and nonconscious behavioral responses to specific contextual cues, distinguishing them from deliberate and conscious behaviors (e.g., Ouellette and Wood 1998, Verplanken and Wood 2006, Wood and Neal 2007). Researchers in economics and marketing, using revealed preference as a central tool, often consider habits as a form of state dependence and view them as optimal choices made by utility-maximizing consumers (e.g., Pollak 1970, Becker and Murphy 1988, Erdem 1996, Dubé et al. 2010).

[2] A notable exception is Target Corporation, which had access to both credit card spending and detailed retail purchase data before selling its credit card portfolio to TD Bank Group in 2012.

[3] Both delinquency and payment variables are available at the month level, at which a month in the data runs from the 14th of a month to the 13th of the following month. Because the billing cycle varies across cardholders, we do not know the exact timing of missed payments. However, as we discuss shortly, this limitation is unlikely to significantly impact our main analysis.

[4] Credit card companies train and deploy a host of scoring algorithms to assess the risk of new applicants and manage existing accounts. These algorithms differ in their objectives, input data types, and outcome variable definitions. In this paper, we focus on a specific type of scoring algorithm that assumes a particular segmentation scheme with the goal of offering a proof of concept on the value of grocery data in making credit decisions. Our chosen segmentation scheme, utilized by our data sponsor, is also widely recognized among industry experts (Dash 2010).

[5] This definition closely corresponds with a key operating performance measure used in the credit card industry known as past due credit card receivables (PD2+). PD2+ represents the percentage of credit card receivables more than 30 days past due (i.e., delinquent at least two months). Our definition of default also aligns with the data sponsor's concept of "serious" delinquency. According to the manager, a significant portion of one-month delinquent cardholders tend to pay off their bills immediately the following month, whereas those who fail to do so within a month have a very low chance of coming back. Our data confirms this insight. Among our final sample consumers, one-month delinquent payers had a 72.4% probability of paying off their debts without entering a two-month delinquency. However, once cardholders entered a two-month delinquency, the probability of repayment dropped to 17.4% (see Online Appendix A.2).

[6] We discuss how this feature affects our interpretation of the results in Section 4.3.

[7] Food items are defined as product groups that belong to either (a) the consumer packaged goods department but not the household items section or (b) the fresh goods department. In our data, there are 162 food items, accounting for approximately 65% of total supermarket revenue.

[8] The results in Table 3 remain largely similar when we run a joint regression of all the food items and the same set of control variables.

[9] To examine which day of the week each segment prefers for grocery shopping, we look at the average of consumer-level trip probabilities across days of the week by segment. We find that all three segments are most likely to shop on weekends. The distribution of trip probabilities across days of the week is more skewed toward weekends for never-delinquents, whereas sloppy payers and defaulters exhibit a relatively flatter distribution.

[10] To compute the variance of grocery expenditures, we use the purchase of items from either the consumer packaged goods department or the fresh goods department. This step minimizes the impact of infrequent purchases of big-ticket items, such as durables, and instead captures the variance of spending on necessities.

[11] When examining the marginal effect of the conditional probability of buying a sale item in a joint regression that incorporates all how-to-buy grocery features, we observe a reversal in the signs of the marginal effect, which may result from correlations between grocery features (see Online Appendix A.3 for more details). Specifically, the average marginal effects for never-delinquents, sloppy payers, and defaulters are −0.542 (0.137), 0.284 (0.117), and 0.258 (0.081), respectively.

[12] The healthiness and convenience of a basket are determined solely by the ratings for the 35 items in Table 3. Three hundred thirty-seven out of 30,089 consumers (1.12%) who did not buy any of the 35 items are assumed to have a grocery basket of median healthiness and convenience. Results are robust when we drop those consumers.

[13] For a review of the modeling approach in the credit scoring literature, see Thomas et al. (2017).

[14] For instance, credit card issuers can offer higher interest rates to sloppy payers to extract more interest revenue or lower credit limits to defaulters to minimize the expected loss of principal.

[15] We evaluate the predictive performance of different hierarchical approaches and a flat approach and find that they have similar performance across all information sets considered.

[16] The imbalanced class distribution could preclude efficient learning, given that 81% of the sample consists of never-delinquent payers, 12% sloppy payers, and 7% defaulters. The concern is particularly salient for a classifier distinguishing between never-delinquent and delinquent payers. To alleviate this concern, we apply different weights to positive and negative classes to balance the distribution during training.

[17] Before partitioning the consumers into 10 folds, we set aside a subset of consumers to be used as a validation set during Platt scaling, which transforms the scores generated by the credit scoring algorithms into probabilities. This process is necessary for simulating credit extension decisions in Section 5. We use stratified sampling to maintain the relative frequency of the three classes (credit card segments) within each fold.

[18] We only observe the sample consumers' credit card payment behaviors with the focal issuer but not with other issuers. To examine how this impacts the estimated predictive value of grocery data, we repeated our analysis using a subset of 19,221 individuals whose usage of the focal credit card is more consistent and frequent. The underlying assumption in this analysis was that data from the focal issuer would more accurately capture the overall credit risk of these individuals compared with those who use the card inconsistently and infrequently. We find that grocery data provides higher incremental out-of-sample accuracy for consistent and frequent card users, which suggests that relying solely on data from the focal issuer may underestimate the predictive value of grocery data.

[19] This type of selection problem is common across a wide range of contexts, such as hiring decisions (e.g., Autor and Scarborough 2008, Li et al. 2020) and bail decisions (e.g., Arnold et al. 2018, Kleinberg et al. 2018).

[20] Implementing such an experiment would require significant administrative costs (e.g., cost of advertising the opportunity and processing applications), monetary costs associated with increased default risk (e.g., loss of principal, debt collection costs), reputational costs (e.g., lower credit ratings), and opportunity costs (e.g., cost of foregoing the opportunity to extend credit to less risky applicants).

[21] Lenders in various markets are subject to responsible lending laws. In the United States, the Equal Credit Opportunity Act mandates lenders to provide consumers with explanations for any adverse actions, such as denying credit. Similarly, the European Union's Consumer Credit Directive and the United Kingdom's Consumer Credit Act oblige lenders to assess consumers' creditworthiness before granting credit and ensure that the credit is suitable for the consumer's financial situation.

[22] Model-based inference techniques include extrapolation, augmentation, iterative reclassification, and parceling. Some use default outcomes of the rejected in other domains as a proxy for their credit risk in the focal domain (e.g., Blattner and Nelson 2021).

[23] We thank the anonymous review team for their constructive suggestions on the simulations in Sections 5 and 6.

[24] Alternatively, one could write down a model in which the lender's decision causally affects the default probability. One example is a model in which the decision variable is credit terms, such as interest rates and credit limits, which can subsequently influence the approved applicants' credit card usage and payment behaviors and, ultimately, their default probabilities.

[25] In practice, the cost of false negatives may not be zero. One reason for this is the possibility of pushback from rejected applicants who might have been approved had lenders not taken grocery data into account.

[26] Whereas we use expected payoffs in Equation (4) as our metric for decision making, our results remain robust when using an alternative metric: predicted segment probabilities (p_i's) without considering payoffs (π_i's).

References

Acolin A, Bricker J, Calem P, Wachter S (2016) Borrowing constraints and homeownership. *Amer. Econom. Rev.* 106(5):625–629.

Ailawadi KL, Neslin SA, Gedenk K (2001) Pursuing the value-conscious consumer: Store brands vs. national brand promotions. *J. Marketing* 65(1):71–89.

Arnold D, Dobbie W, Yang CS (2018) Racial bias in bail decisions. *Quart. J. Econom.* 133(4):1885–1932.

Autor DH, Scarborough D (2008) Does job testing harm minority workers? Evidence from retail establishments. *Quart. J. Econom.* 123(1):219–277.

Bajari P, Chernozhukov V, Hortaçsu A, Suzuki J (2019) The impact of big data on firm performance: An empirical investigation. *AEA Papers Proc.* 109:33–37.

Ball I (2019) Scoring strategic agents. Preprint, submitted September 4, https://arxiv.org/abs/1909.01888.

Barakova I, Bostic RW, Calem PS, Wachter SM (2003) Does credit quality matter for homeownership? *J. Housing Econom.* 12(4):318–336.

Becker GS, Murphy KM (1988) A theory of rational addiction. *J. Political Econom.* 96(4):675–700.

Bell DR, Corsten D, Knox G (2011) From point of purchase to path to purchase: How preshopping factors drive unplanned buying. *J. Marketing* 75(1):31–45.

Berg T, Fuster A, Puri M (2022) Fintech lending. *Annual Rev. Financial Econom.* 14:187–207.

Berg T, Burg V, Gombović A, Puri M (2020) On the rise of fintechs: Credit scoring using digital footprints. *Rev. Financial Stud.* 33(7):2845–2897.

Berger PG, Ofek E (1995) Diversification's effect on firm value. *J. Financial Econom.* 37(1):39–65.

Björkegren D, Grissen D (2020) Behavior revealed in mobile phone usage predicts credit repayment. *World Bank Econom. Rev.* 34(3):618–634.

Björkegren D, Blumenstock JE, Knight S (2020) Manipulation-proof machine learning. Preprint, submitted April 8, https://arxiv.org/abs/2004.03865.

Blattner L, Nelson S (2021) How costly is noise? Data and disparities in consumer credit. Preprint, submitted May 17, https://arxiv.org/abs/2105.07554.

Cawley GC, Talbot NL (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Machine Learn. Res.* 11:2079–2107.

Chan T, Hamdi N, Hui X, Jiang Z (2022) The value of verified employment data for consumer lending: Evidence from Equifax. *Marketing Sci.* 41(4):795–814.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R, eds. *Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 785–794.

Collins T (2021) Keyna and Ghana lead world in mobile money. *African Business Online* (July 26), https://african.business/2021/07/technology-information/kenya-and-ghana-lead-world-in-mobile-money.

Consumer Financial Protection Bureau (2016) Who are the credit invisibles? How to help people with limited credit histories. Reports, Consumer Financial Protection Bureau, Washington, DC.

Dash E (2010) Risky borrowers find credit available again, at a price. *CNBC Online* (December 13), https://www.cnbc.com/2010/12/13/risky-borrowers-find-credit-available-again-at-a-price.html.

De Cnudde S, Moeyersoms J, Stankova M, Tobback E, Javaly V, Martens D (2019) What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance. *J. Oper. Res. Soc.* 70(3):353–363.

Del Boca D, Lusardi A (2003) Credit market constraints and labor market decisions. *Labour Econom.* 10(6):681–703.

Demirguc-Kunt A, Klapper L, Singer D, Ansar S (2018) *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution* (World Bank Publications, Washington, DC).

Demirgüç-Kunt A, Klapper L, Singer D, Ansar S (2022) *The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19* (World Bank Publications, Washington, DC).

Di Maggio M, Ratnadiwakara D, Carmichael D (2022) Invisible primes: Fintech lending with alternative data. NBER Working Paper No. 29840, National Bureau of Economic Research, Cambridge, MA.

Dickson PR, Sawyer AG (1990) The price knowledge and search of supermarket shoppers. *J. Marketing* 54(3):42–53.

Dorfleitner G, Priberny C, Schuster S, Stoiber J, Weber M, de Castro I, Kammler J (2016) Description-text related soft information in peer-to-peer lending—Evidence from two leading European platforms. *J. Banking Finance* 64:169–187.

Dubé J-P, Hitsch GJ, Rossi PE (2010) State dependence and alternative explanations for consumer inertia. *RAND J. Econom.* 41(3):417–445.

Erdem T (1996) A dynamic analysis of market structure based on panel data. *Marketing Sci.* 15(4):359–378.

Evans DS, Jovanovic B (1989) An estimated model of entrepreneurial choice under liquidity constraints. *J. Political Econom.* 97(4): 808–827.

Fannie Mae (2021) Fannie Mae introduces new underwriting innovation to help more renters become homeowners. Accessed June 26, 2024, https://www.fanniemae.com/newsroom/fannie-mae-news/fannie-mae-introduces-new-underwriting-innovation-help-more-renters-become-homeowners.

Farboodi M, Veldkamp L (2021) A model of the data economy. NBER Working Paper No. 28427, National Bureau of Economic Research, Cambridge, MA.

Farrell J, Klemperer P (2007) Chapter 31: Coordination and lock-in: Competition with switching costs and network effects. Armstrong M, Porter R, eds. *Handbook of Industrial Organization*, vol. 3 (Elsevier, Amsterdam), 1967–2072.

FICO (2022) Introducing the ultrafico score. Accessed June 26, 2024, https://www.fico.com/ultrafico.

Frankel A, Kartik N (2022) Improving information from manipulable data. *J. Eur. Econom. Assoc.* 20(1):79–115.

Gordon BR, Sun B (2015) A dynamic model of rational addiction: Evaluating cigarette taxes. *Marketing Sci.* 34(3):452–470.

Guadagni PM, Little JD (1983) A logit model of brand choice calibrated on scanner data. *Marketing Sci.* 2(3):203–238.

Guidotti R, Coscia M, Pedreschi D, Pennacchioli D (2015) Behavioral entropy and profitability in retail. *2015 IEEE Internat. Conf. Data Sci. Adv. Anal.* (IEEE, Piscataway, NJ), 1–10.

Heckman JJ (1981) Chapter 3: Statistical models for discrete panel data. Manski C, McFadden D, eds. *Structural Analysis of Discrete Data with Econometric Applications* (MIT Press, Cambridge, MA), 114–178.

Hu L, Immorlica N, Vaughan JW (2019) The disparate effects of strategic manipulation. Chouldechova A, Diaz F, eds. *Proc. Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 259–268.

Hughes-Cromwick E, Coronado J (2019) The value of US government data to US business decisions. *J. Econom. Perspect.* 33(1): 131–146.

Iyer R, Khwaja AI, Luttmer EF, Shue K (2016) Screening peers softly: Inferring the quality of small borrowers. *Management Sci.* 62(6):1554–1577.

Jensen MC (1986) Agency costs of free cash flow, corporate finance, and takeovers. *Amer. Econom. Rev.* 76(2):323–329.

Jones CI, Tonetti C (2020) Nonrivalry and the economics of data. *Amer. Econom. Rev.* 110(9):2819–2858.

Khare A, Inman JJ (2006) Habitual behavior in American eating patterns: The role of meal occasions. *J. Consumer Res.* 32(4):567–575.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Quart. J. Econom.* 133(1):237–293.

Lee L, Ariely D (2006) Shopping goals, goal concreteness, and conditional promotions. *J. Consumer Res.* 33(1):60–70.

Li D, Raymond LR, Bergman P (2020) Hiring as exploration. NBER Working Paper No. 27736, National Bureau of Economic Research, Cambridge, MA.

McAlister L (1982) A dynamic attribute satiation model of variety-seeking behavior. *J. Consum. Res.* 9(2):141–150.

McGurran B (2023) How utility bills can boost your credit score. *Experian Online* (August 16), https://www.experian.com/blogs/ask-experian/does-paying-utility-bills-help-your-credit-score.

Menon S, Kahn BE (1995) The impact of context on variety seeking in product choices. *J. Consumer Res.* 22(3):285–295.

Myerson RB (1982) Optimal coordination mechanisms in generalized principal–agent problems. *J. Math. Econom.* 10(1):67–81.

Netzer O, Lemaire A, Herzenstein M (2019) When words sweat: Identifying signals for loan default in the text of loan applications. *J. Marketing Res.* 56(6):960–980.

Osborne M (2011) Consumer learning, switching costs, and heterogeneity: A structural examination. *Quant. Marketing Econom.* 9(1):25–70.

Óskarsdóttir M, Bravo C, Sarraute C, Vanthienen J, Baesens B (2019) The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Appl. Soft Comput.* 74:26–39.

Ouellette JA, Wood W (1998) Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psych. Bull.* 124(1):54–74.

Pollak RA (1970) Habit formation and dynamic demand functions. *J. Political Econom.* 78(4 part 1):745–763.

Ratner RK, Kahn BE, Kahneman D (1999) Choosing less-preferred experiences for the sake of variety. *J. Consumer Res.* 26(1):1–15.

San Pedro J, Proserpio D, Oliver N (2015) Mobiscore: Toward universal credit scoring from mobile phone data. Ricci F, Bontcheva K, Conlan O, Lawless S, eds. *Internat. Conf. User Model. Adaptation Personalization* (Springer International Publishing, Cham, Switzerland), 195–207.

Shannon CE (1948) A mathematical theory of communication. *Bell Systems Tech. J.* 27(3):379–423.

Shevlin R (2019) Can Apple or Amazon become full stack banks? *Forbes Online* (September 3), https://www.forbes.com/sites/ronshevlin/2019/09/03/can-apple-or-amazon-become-full-stack-banks.

Solis A (2017) Credit access and college enrollment. *J. Political Econom.* 125(2):562–622.

Thomas L, Crook J, Edelman D (2017) *Credit Scoring and Its Applications*, 2nd ed. (SIAM, Philadelphia).

Trijp HCV, Hoyer WD, Inman JJ (1996) Why switch? Product category–level explanations for true variety-seeking behavior. *J. Marketing Res.* 33(3):281–292.

Urbany JE, Dickson PR, Kalapurakal R (1996) Price search in the retail grocery market. *J. Marketing* 60(2):91–104.

Verhoeven AA, Adriaanse MA, Evers C, de Ridder DT (2012) The power of habits: Unhealthy snacking behaviour is primarily predicted by habit strength. *British J. Health Psych.* 17(4):758–770.

Verplanken B, Wood W (2006) Interventions to break and create consumer habits. *J. Public Policy Marketing* 25(1):90–103.

Vissing-Jorgensen A (2021) Consumer credit: Learning your customer's default risk from what (s)he buys. Working paper, Haas School of Business, University of California Berkeley, Berkeley, CA, NBER, Cambridge, MA and CEPR, London, UK.

Walters RG (1991) Assessing the impact of retail price promotions on product substitution, complementary purchase, and interstore sales displacement. *J. Marketing* 55(2):17–28.

Wood W, Neal DT (2007) A new look at habits and the habit-goal interface. *Psych. Rev.* 114(4):843–863.