On the Rise of FinTechs: Credit Scoring Using Digital Footprints

Tobias Berg, Valentin Burg, Ana Gombović, Manju Puri

The Review of Financial Studies, 2020.

叶鑫 2021/05/27

Background & Motivation

- The growth of the internet leaves a trace of **simple, easily accessible** information about almost every individual worldwide – a trace that we label “digital footprint”.
- A key reason for the existence of financial intermediaries is their superior ability to access and process information relevant for screening and monitoring of borrowers.
- If digital footprints yield significant information on predicting defaults then FinTechs – with their superior ability to access and process digital footprints – can threaten the information advantage of financial intermediaries and thereby challenge financial intermediaries’ business models

Research question

1. Whether the digital footprint helps augment information traditionally considered to be important for default prediction?

2. Whether it can be used for the prediction of consumer payment behavior and defaults?

Research Data: Digital Variable

- Our data set contains a set of ten digital footprint variables: the device type (for example, tablet or mobile), the operating system (for example, iOS or Android), etc.
- Our data set also contains a credit score from a private credit bureau.

Digital footprint variables		
Device type	Device type. Main examples: Desktop, Tablet, Mobile.	Categorical variable
Operating system	Operating system. Main examples: Windows, iOS, Android, Macintosh	Categorical variable
E-mail host	E-mail host. Main examples: Gmx, Web, T-Online, Gmail, Yahoo, Hotmail	Categorical variable
Channel	Channel through which customer comes to Web site. Main examples: Paid (including paid and retargeted clicks), Direct, Affiliate, Organic	Categorical variable
Checkout time	Time of day of purchase	Numerical variable (0–24 hr)
Do-not-track setting	Dummy equal to one if customer does not allow tracking of device and operating system information, and channel	Dummy variable
Name in E-mail	Dummy equal to one if first or last name of customer is part of e-mail address	Dummy variable
Number in E-mail	Dummy equal to one if a number is part of e-mail address	Dummy variable
Is lowercase	Dummy equal to one if first name, last name, street, or city are written in lowercase	Dummy variable
E-mail error	Dummy equal to one if e-mail address contains an error in the first trial (Note: Clients can only order if they register with a correct e-mail address)	Dummy variable

Research Contents

- We first provides default rates by credit bureau score quintile and default rates by category of each of the digital footprint variables.

- We provide a more formal analysis of the discriminatory power of digital footprint variables by constructing receiver operating characteristics and determining the area under the curve (AUC).

- Our results are robust to a large set of robustness tests, like time or region fixed effects, age, or gender, and results are robust to various default definitions and sample splits and hold out-of-sample as well.

- Fourth and finally, we discuss implications of our findings for the behavior of consumers, firms and regulators.

Research Conclusion

- Our results suggest that even the simple, easily accessible variables from the digital footprint proxy for income, character, and reputation and are highly valuable for default prediction.

- Second, we document that default rates drop significantly after the introduction of the digital footprint, thereby highlighting the economic benefit to the E-commerce firm of using the digital footprint.

- Third, we show that digital footprints work equally well for unscorable as for scorable customers, so that has the potential to boost financial inclusion for the 2 billion adults worldwide that lack access to credit.

- Furthermore, we show that digital footprints today can forecast future changes in the credit bureau score.

Research Innovation

- Prior papers have highlighted the role of relationship-specific information for lending as well as the informativeness of nontraditional data sources.

- Our paper differs from the prior literature in that the information we are looking at is provided simply by accessing or registering on a Web site and, therefore, stands out for their ease of collection.

- Our results imply that barriers to entry in financial intermediation might be lower in a digital world, and the digital footprint can be used to process applications faster than traditional lenders.

- A credit score based on the digital footprint should therefore serve as a benchmark for other models that use more elaborate sources of information

Research Data: Sample

- **Sample:** We access data about 270,399 purchases(>100 €) by invoice from an E-commerce company selling furniture in Germany between October 2015 and December 2016.

- **Dependent variable(default):** A customer who does not pay after three reminders is in default, and the claim is transferred to a debt collection agency, on average 3.5 months from the order date.

- **The credit bureau score:** draws on credit history data from various banks,sociodemographic data, and payment behavior data sourced from retail sales firms, telecommunication companies, and utilities, which is requested for purchases exceeding EUR 100.

- **Scorable customers:** We label those customers for whom a credit bureau score exists “scorable customers.”

Empirical result: Descriptive statistics

- The credit bureau score is available for **94%** of the sample, **0.9%** default rate and unavailable is **6%** of the sample, **2.5%** default rate.
- Descriptive statistics for the sample without credit bureau score are similar with respect to order amount and gender, with **age** being somewhat lower, consistent with the idea that it takes time to build up a credit history.

Table 1
Descriptive statistics

A. Customers with credit bureau score

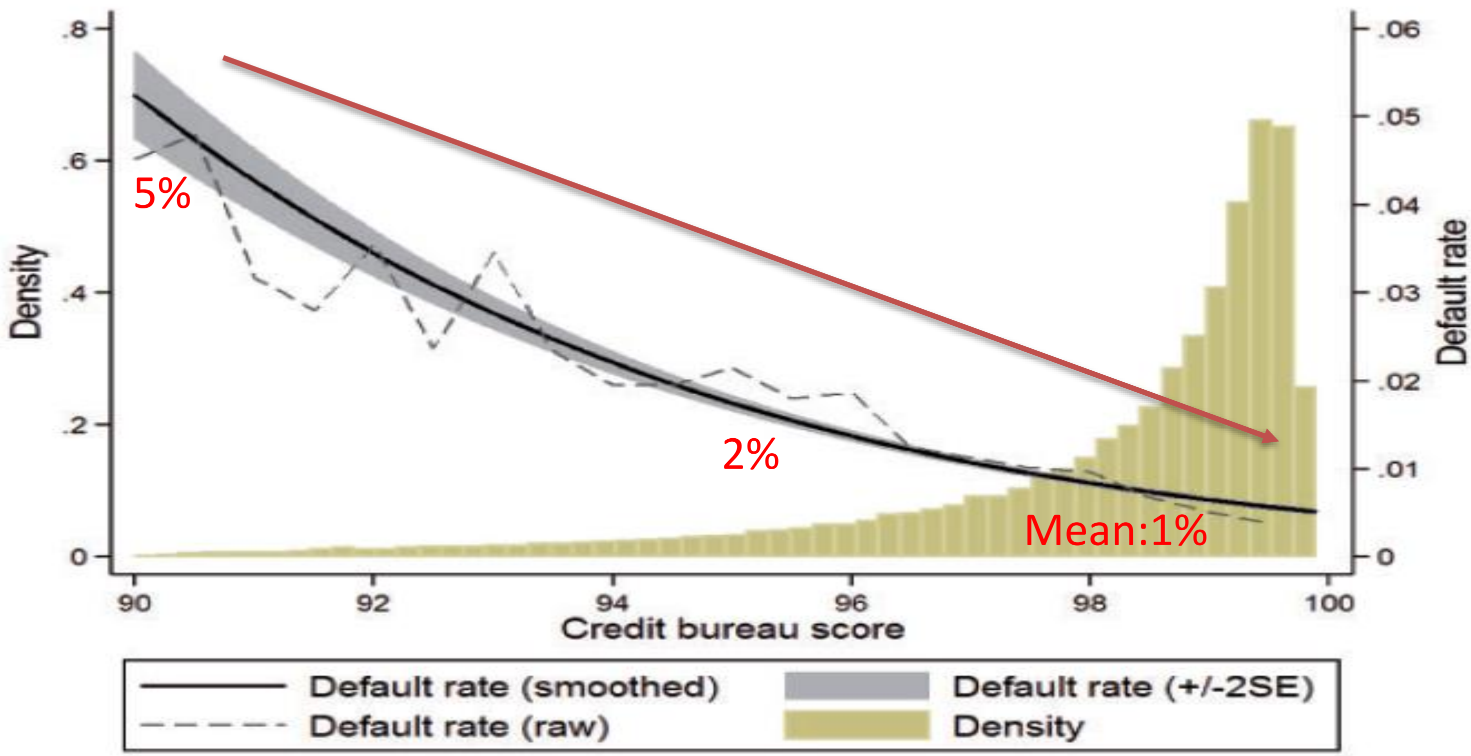
Variable	Unit	N	Mean	SD	P25	Median	P75
Order and customer							
Order amount	Euro	254,819	317.75	317.10	119.99	218.90	399.98
Gender	Dummy (0=male, 1=female)	254,819	0.66	0.47	0	1	1
Age[a]	Number	254,613	45.06	13.31	34	45	54
Credit bureau score	Number (0=worst, 100=best)	254,819	98.11	2.05	97.58	98.86	99.41
Payment behavior							
Default	Dummy (0/1)	254,819	0.009	0.096	0	0	0

B. Customers without credit bureau score

Variable	Unit	N	Mean	Std.	P25	Median	P75
Order and customer							
Order amount	Euro	15,580	324.57	319.22	119.99	221.60	399.99
Gender	Dummy (0=male, 1=female)	15,580	0.70	0.46	0	1	1
Age[a]	Number	555	38.20	10.46	30	35	46
Credit bureau score	Number (0=worst, 100=best)	15,580	na	na	na	na	na
Payment behavior							
Default	Dummy (0/1)	15,580	0.025	0.156	0	0	0

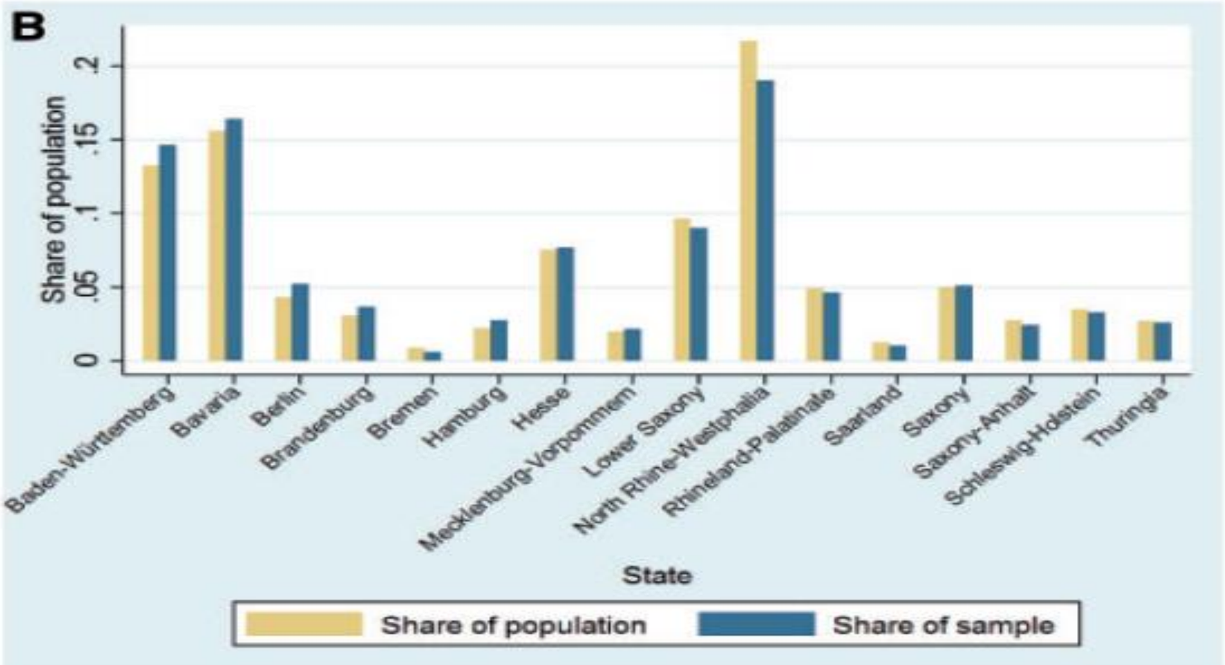
Empirical result: Descriptive statistics

- Default rates grow **exponentially** when credit bureau scores decrease, with a credit bureau of 95 corresponding to a 2% default rate and a credit bureau of 90 corresponding to a 5% default rate.
- Standard errors are generally higher for lower credit bureau scores (due to the smaller number of observations), but do not exceed 0.25% even for a credit bureau score as low as 90.



Empirical result: Representativeness of data set

- Our data set is largely representative of the geographic distribution of the German population overall.
- The mean customer age is 45.06 years, comparable both to the mean age of 43.77 in the German population.
- The average default rate in our sample is 1.0%, implying a scaled-up annualized default rate of 3.0% (these default window of approximately 4 months), which is consistent with the major German credit bureau reports(2.5%).
- Taken together, this evidence suggests that default rates in our sample are largely representative of a typical consumer loan sample in Germany.



Empirical result: Univariate results

Table 2
Credit bureau score, digital footprint variables, and default rates (scorable customers)

Variable	Value	Observations	Proportion (%)	Default rate (%)	*t*-test against baseline
Credit bureau score	All	254,819	100	0.94	
(by quintile)	Q1 - lowest	50,980	20	2.12	Baseline
	Q2	50,949	20	1.02***	(−14.17)
	Q3	50,991	20	0.68***	(−19.51)
	Q4	51,181	20	0.47***	(−23.37)
	Q5 - highest	50,718	20	0.39***	(−24.89)
Device	All	254,819	100	0.94	
	Desktop	145,879	57	0.74	Baseline
	Tablet	45,575	18	0.91***	(3.62)
	Mobile	26,808	11	2.14***	(21.84)
	Do-not-track setting	36,557	14	0.88***	(2.90)
Operating system	All	254,819	100	0.94	
	Windows	124,605	49	0.74	Baseline
	iOS	41,478	16	1.07***	(6.35)
	Android	29,089	11	1.79***	(16.64)
	Macintosh	21,163	8	0.69	(−0.79)
	Other	1,927	1	1.09*	(1.74)
	Do-not-track setting	36,557	14	0.88***	(2.66)
E-mail host	All	254,819	100	0.94	
	Gmx (partly paid)	58,609	23	0.82	Baseline
	Web (partly paid)	54,867	22	0.86	(0.70)
	T-Online (affluent customers)	30,279	12	0.51***	(−5.32)
	Gmail (free)	27,845	11	1.25***	(6.02)
	Yahoo (free, older service)	11,923	5	1.96***	(11.33)
	Hotmail (free, older service)	10,241	4	1.45***	(6.11)
	Other	61,055	24	0.90	(1.38)

Empirical result: Univariate results

Table 2
Credit bureau score, digital footprint variables, and default rates (scorable customers)

Variable	Value	Observations	Proportion (%)	Default rate (%)	t-test against baseline
Channel	All	254,819	100	0.94	
	Paid	111,399	44	1.11	Baseline
	Direct	45,183	18	0.84***	(−4.78)
	Affiliate	24,770	10	0.64***	(−6.68)
	Organic	18,295	7	0.86***	(−3.00)
	Other	18,615	7	0.69***	(−5.24)
	Do-not-track setting	36,557	14	0.88***	(−3.69)
Checkout time	All	254,819	100	0.94	
	Evening (6 p.m.-midnight)	108,549	43	0.85	Baseline
	Night (midnight-6 a.m.)	6,913	3	1.97***	(9.49)
	Morning (6 a.m.-noon)	46,601	18	1.09***	(4.55)
	Afternoon (noon-6 p.m.)	92,756	36	0.89	(0.91)
Do-not-track setting	All	254,819	100	0.94	
	No	218,262	86	0.94	Baseline
	Yes	36,557	14	0.88	(−1.12)
Name in e-mail	All	254,819	100	0.94	
	No	71,017	28	1.24	Baseline
	Yes	183,802	72	0.82***	(−9.99)
Number in e-mail	All	254,819	100	0.94	
	No	213,649	84	0.84	Baseline
	Yes	41,170	16	1.41***	(10.95)
Is lowercase	All	254,819	100	0.94	
	No	235,569	92	0.84	Baseline
	Yes	19,250	8	2.14***	(18.07)
E-mail error	All	254,819	100	0.94	
	No	251,319	99	0.88	Baseline
	Yes	3,500	1	5.09***	(25.71)

Empirical result: Combination of digital footprint variables

- As most of the digital footprint variables are categorical variables, We therefore report Cramér's V.
- The Cramér's V between the credit bureau score and the digital footprint variables is economically small, with values ranging between 0.01 and 0.07. This suggests that digital footprint variables act as complements rather than substitutes for credit bureau scores, a claim we will analyze more formally below in a multivariate regression setup.

Table 3
Correlation/association between credit bureau score, digital footprint, and control variables (scorable customers)

	Credit bureau score	Device type	Operating system	E-mail host	Channel	Checkout time	Name in e-mail	Number in e-mail	Is lowercase	E-mail error
Main variables										
Credit bureau score[a]	1.00***	0.07***	0.05***	0.07***	0.03***	0.03***	0.01***	0.07***	0.02***	0.00
Device type		1.00***	0.71***[b]	0.07***	0.06***[b]	0.04***	0.05***	0.06***	0.07***	0.01***
Operating system			1.00***	0.08***	0.06***[b]	0.04***	0.06***	0.08***	0.06***	0.01***
E-mail host				1.00***	0.03***	0.03***	0.08***	0.18***	0.04***	0.06***
Channel					1.00***	0.02***	0.01***	0.02***	0.04***	0.02***
Checkout time[a]						1.00***	0.01***	0.01***	0.01***	0.01*
Name in e-mail							1.00***	0.22***	0.01***	0.02***
Number in e-mail								1.00***	0.02***	0.00**
Is lowercase									1.00***	0.03***
E-mail error										1.00***

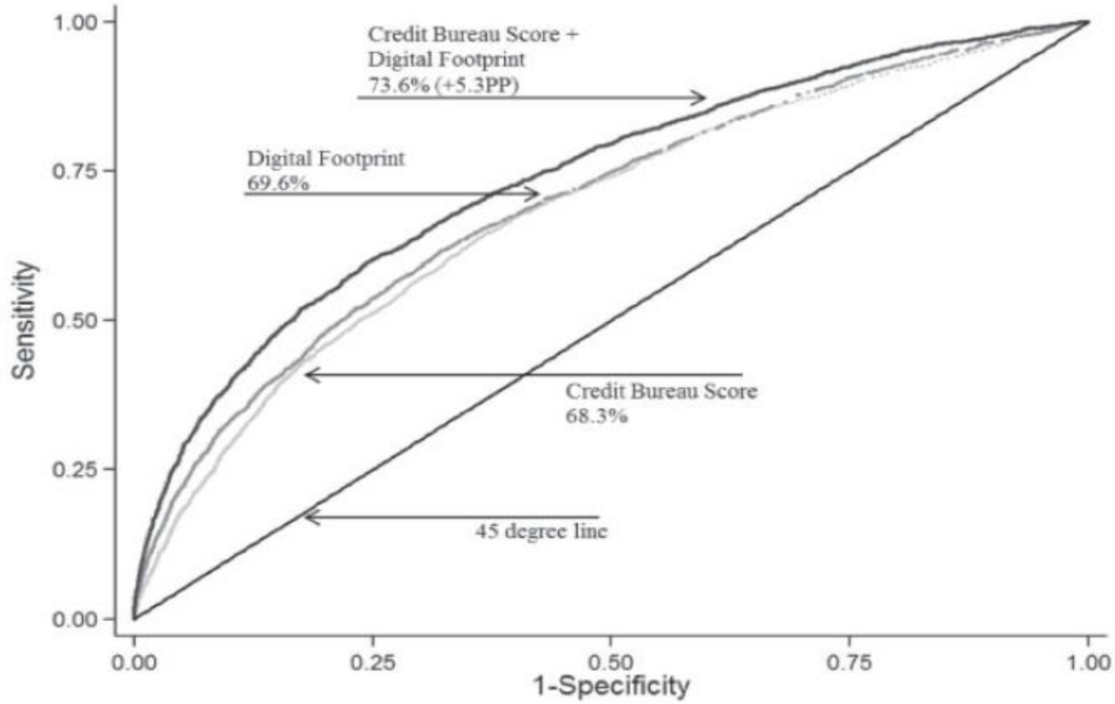
Multivariate results: Digital footprint and default

- We use a logistic regression and report the AUC for every specification.
- For categorical variables, all coefficients need to be interpreted relative to the baseline level. We always choose the most popular category in a variable as the baseline level.

	(1) Credit bureau bureau score		(2) Digital footprint		(3) Credit bureau score & digital footprint		(4) Credit bureau score & digital footprint, further controls	
Variables	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat
Credit bureau score	−0.17***	(−7.89)			−0.15***	(−6.67)	−0.14***	(−5.90)
Device type & operating system[a]								
Desktop/Windows			Baseline		Baseline		Baseline	
Desktop/Macintosh			−0.07	(−0.53)	−0.13	(−1.03)	−0.19	(−1.52)
Tablet/Android			0.29***	(3.19)	0.29***	(3.06)	0.33***	(3.44)
Tablet/iOS			0.08	(1.05)	0.08	(0.97)	0.07	(0.89)
Mobile/Android			1.05***	(17.25)	0.95***	(15.34)	1.01***	(16.13)
Mobile/iOS			0.72***	(9.07)	0.57***	(6.73)	0.61***	(7.26)
Checkout time								
Evening (6 p.m.-midnight)			Baseline		Baseline		Baseline	
Morning (6 a.m.-noon)			0.28***	(4.50)	0.28***	(4.60)	0.29***	(4.75)
Afternoon (noon-6 p.m.)			0.08	(1.42)	0.08	(1.47)	0.10*	(1.92)
Night (midnight-6 a.m.)			0.79***	(7.73)	0.75***	(7.09)	0.72***	(6.68)
Do-not-track setting			−0.02	(−0.25)	−0.07	(−0.91)	−0.09	(−1.19)
Name in e-mail			−0.28***	(−5.67)	−0.29***	(−5.70)	−0.29***	(−5.59)
Number in e-mail			0.26***	(4.50)	0.23***	(3.91)	0.22***	(3.85)
Is lowercase			0.76***	(13.10)	0.74***	(13.20)	0.74***	(13.24)
E-mail error			1.66***	(20.00)	1.67***	(20.36)	1.70***	(20.37)
Constant	12.42***	(5.76)	−4.92***	(−62.87)	9.97***	(4.48)	9.04***	(4.06)
AUC	**0.683**		**0.696**		**0.736**		**0.762**	
(SE)	**(0.006)**		**(0.006)**		**(0.005)**		**(0.005)**	

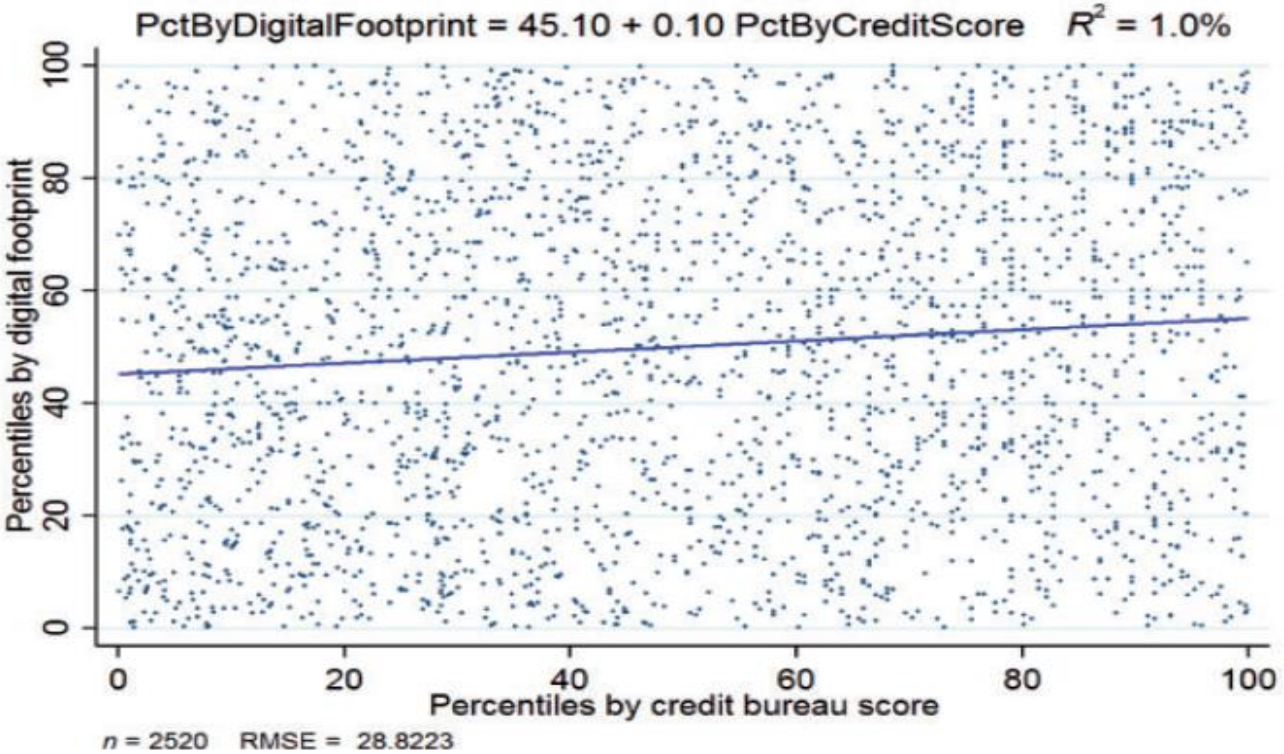
Multivariate results: Digital footprint and default

- Interestingly, digital footprint variables have an AUC of 69.6%, which is higher than the AUC of the credit bureau score.
- These results suggest that even simple, easily accessible variables from the digital footprint are as useful in predicting defaults as the credit bureau score.
- The AUC of the combined model (73.6%) is significantly higher than the AUC of each of the stand-alone models



Multivariate results: Digital footprint and default

- We construct a default prediction using only the digital footprint variables for each observation in our sample.
- Figure 4 clearly shows that the correlation between credit bureau score and digital footprint is very low (R2 of 1.0%, implying a correlation of approximately 10%).
- These results confirm our prior observation that the digital footprint acts as a complement, rather than a substitute, of the credit bureau score.



Empirical result: Out-of-sample tests

- For the out-of-sample tests, we use Nx2-fold cross-validation.
- Reassuringly, the OOS-OOT AUC is very similar to both the in-sample and the out-of-sample AUC.
- In particular, there seems to be little evidence that the link between digital footprints and defaults changes quickly over time.

Out-of-sample estimates

	(1) Baseline (in-sample)	(2) Out-of-sample	(3) Out-of-sample/out-of-time
AUC credit bureau score	0.683	0.681	0.691
N	254,819	254,819	74,543
AUC digital footprint	0.696	0.688	0.692
N	254,819	254,819	74,543
AUC credit bureau score + Digital footprint	0.736	0.728	0.739
N	254,819	254,819	74,543
AUC credit bureau score + Digital footprint, fixed effects	0.762	0.734	0.730
N	254,613	254,613	74,543

Empirical result: Alternative default definitions and sample splits

Overall, the robustness tests suggest that digital footprints predict default as well or even better than the credit bureau score, and digital footprint and credit bureau score are complements rather than substitutes—is robust for different default definitions and various sample splits.

Robustness tests (scorable customers)

A. Default definition	(1)	(2)	(3)	(4)
	Baseline (default = transfer to collection agency)	Default = Write-down	Exclude cases of fraud (9% of defaults)	Loss given default (R^2 reported)
AUC credit bureau score	0.683	0.692	0.681	0.013
AUC Digital footprint	0.696	0.723	0.691	0.062
AUC credit bureau score + digital footprint	$\boxed{0.736}$	$\boxed{0.757}$	$\boxed{0.730}$	$\boxed{0.069}$
N	254,819	254,819	254,604	2,384

B. Sample splits	(1)	(2)	(3)	(4)
	Small orders < EUR 218.91	Large orders ≥ EUR 218.91	Female	Male
AUC credit bureau score	0.688	0.678	0.689	0.670
AUC Digital footprint	0.711	0.689	0.697	0.700
AUC credit bureau score + digital footprint	$\boxed{0.749}$	$\boxed{0.729}$	$\boxed{0.743}$	$\boxed{0.724}$
N	127,410	127,409	168,374	86,445

Empirical result: External validity

- In particular, we test whether digital footprints today can forecast future changes in the credit bureau score.

- If a good digital footprint today predicts an increase in the credit bureau score in the future, then this is evidence that digital footprints matter for other loan products as well.

- We therefore run regressions of the form:

$$\Delta(\text{ Credit Score }_{t+1}, \text{Credit Score }_{t}) = \beta_0 + \beta_1 \Delta(DF_t, \text{ Credit Score }_{t}) + X + \varepsilon$$

Empirical result: External validity

- Taken together, the evidence suggests that digital footprints today forecast subsequent changes in credit bureau scores. This result provides a window into the traditional banking world.
- As credit bureau scores are known to predict default rates for traditional loan products, our results point to the usefulness of digital footprints for traditional loan products as well.

Dependent variable	(1) Δ ($CreditScore_{t+1}$, $CreditScore_t$)	(2) Δ ($CreditScore_{t+1}$, $CreditScore_t$)	(3) Δ ($CreditScore_{t+1}$, $CreditScore_t$)	(4) Δ ($CreditScore_{t+1}$, $CreditScore_t$)
Δ ($DigitalFootprint_t$, $CreditBureauScore_t$)	−75.86***	−28.43***	−30.11***	
	(−11.86)	(−4.64)	(−5.05)	
Q1 (-100% to -0.49%)				0.40**
				(2.52)
Q2 (-0.49% to -0.25%)				0.15*
				(1.75)
Q3 (-0.25% to -0.05%)				baseline
Q4 (-0.05% to +0.35%)				0.08
Q4 (-0.05% to +0.35%)				(0.91)
Q5 (+0.35% to +100%)				−0.39***
				(−3.04)
$CreditBureauScore_t$		−0.43***	−0.42***	−0.42***
		(−13.47)	(−13.28)	(−10.05)
Constant	0.37***	41.99***		
	(8.75)	(13.51)	absorbed	absorbed
Month & region fixed effects	No	No	Yes	Yes
Observations	17,646	17,646	17,646	17,646
Adj. R^2	.028	.071	.081	.081

Economic mechanism

- We cannot fully decompose the informativeness of the digital footprint into one part that proxies for financial characteristics and another part that proxies for what is traditionally viewed as soft information.
- We decompose the overall informational content of the digital footprint into each of the individual variables.

Marginal AUC for digital footprint variables and combinations of digital footprint variables

A. Individual digital footprint variables (dependent variable: default (0/1))

Variable	Stand-alone AUC (%)	Marginal AUC (PP)
Computer & operating system	59.03	+1.71***
E-mail host	59.78	+2.44***
E-mail Host: paid versus nonpaid dummy	53.80	+0.98***
E-mail Host: Variation within nonpaid e-mail hosts	57.82	+1.79***
Channel	54.95	+0.70***
Checkout time	53.56	+0.63***
Do not track setting	50.40	+0.14*
Name in e-mail	54.61	+0.30**
Number in e-mail	54.15	+0.19**
Is lowercase	54.91	+1.15***
E-mail error	53.08	+1.78***

B. Combinations of digital footprint variables (dependent variable: default (0/1))

Economic mechanism

- The first row of panel B categorizes digital footprint variables by their financial costs to switch from one to another， providing suggestive evidence that digital footprints contain information over and above purely financial characteristics.
- We see that both variables determined by a single action and variables determined during each purchase process anew significantly contribute to the informativeness of the digital footprint.

B. Combinations of digital footprint variables (dependent variable: default (0/1))

Variables	Stand-alone AUC (%)	Marginal AUC (PP)
Potential proxy for income		
Potential proxy for income, financially costly to change (computer & operating system, e-mail host: paid vs. nonpaid dummy)	61.03	+2.20
Unlikely to be a proxy for income, not financially costly to change (nonpaid e-mail host, channel, checkout time, do not track setting, name in e-mail, number in e-mail, is lowercase, e-mail error)	67.35	+8.52
Impact on everyday behavior		
Requires one-time action only (computer & operating system, e-mail host, do not track setting, name in e-mail, number in e-mail)	64.92	+7.25
Requires thinking about how to behave during every individual purchase (channel, checkout time, is lowercase, e-mail error)	62.30	+4.63

Access to credit for the unbanked

- Especially in developing countries, the inability of people without bank accounts to participate in financial services is usually caused by the lack of information infrastructure (such as credit bureau scores).
- Interestingly, the AUC of the model using the digital footprint only is similar for unscorable customers compared to the AUC for scorable customers (72.2% vs. 69.6%)

Default regressions (unscorable customers)

Variables	(1) Digital footprint for unscorable customers Coef.	z-stat	(2) For comparison: Digital footprint for scorable customers (Column 2 of Table 4) Coef.	z-stat	(3) Digital footprint for unscorable customers, fixed effects Coef.	z-stat
Computer & operating system						
Desktop/Windows	Baseline		Baseline		Baseline	
Desktop/Macintosh	−0.26	(−1.10)	−0.07	(−0.53)	−0.26	(−1.06)
Tablet/Android	−0.22	(−0.86)	0.29***	(3.19)	−0.11	(−0.44)
Tablet/iOS	−0.45*	(−1.72)	0.08	(1.05)	−0.45*	(−1.67)
Mobile/Android	1.07***	(5.97)	1.05***	(17.25)	1.08***	(5.38)
Mobile/iOS	0.63***	(2.69)	0.72***	(9.07)	0.69***	(2.76)
Control for *Gender*, *Item category*, *Loan amount*, and month and region fixed effects	No		No		Yes	
Observations	15,580		254,819		15,580	
Pseudo R^2	.0906		.0524		.1645	
AUC	**0.722**		**0.696**		**0.803**	
(SE)	**(0.014)**		**(0.006)**		**(0.011)**	
Difference to AUC=50%	0.222***		0.196***		0.302***	
AUC (OOS)	**0.684**		**0.688**		**0.659**	

Access to credit for the unbanked

- With the dramatic increase in the number of people with mobile phones in emerging markets, digital footprints are available even in countries with few official and reliable records.
- We therefore argue that digital footprints are unique in their ability to significantly extend access to credit for the unbanked.
- We have the vision to give billions of unbanked people access to credit when credit bureaus scores do not exist, thereby fostering financial inclusion and lowering inequality

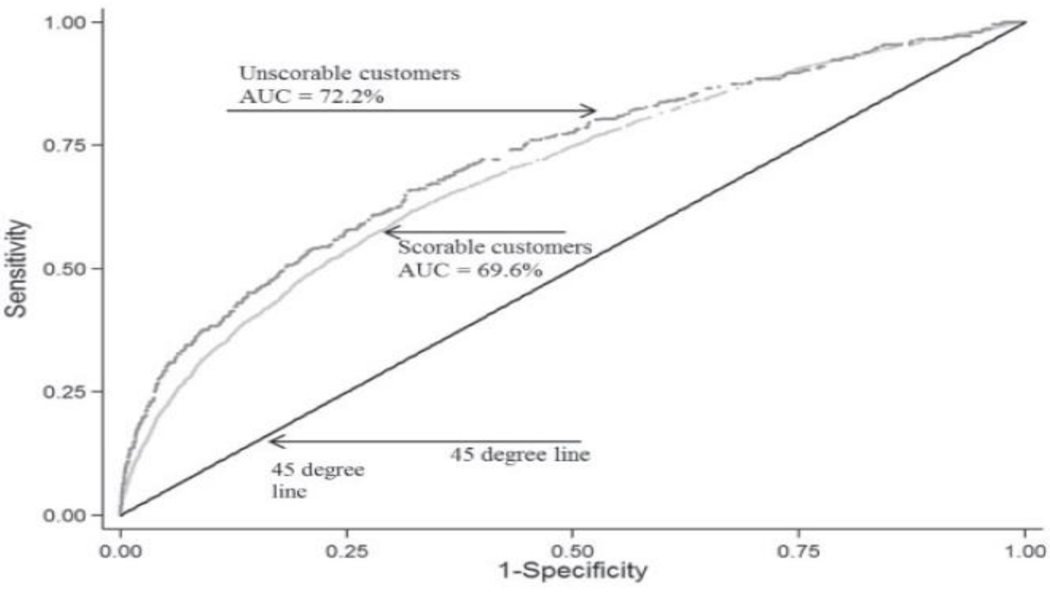


Figure 6
AUC for scorable versus unscorable customers

Conclusion

- We show that even simple, easily accessible variables from the digital footprint match the information content of credit bureau scores. Furthermore, digital footprints complement rather than substitute for credit bureau information.

- We document that default rates drop significantly after adoption of the digital footprint, and customers with good digital footprints gain access to credit while customers with poor digital footprints lose access to credit.

- We also show that the discriminatory power for unscorable customers matches the discriminatory power for scorable customers.

- Given the widespread adaption of smartphones and corresponding digital footprints. The use of digital footprints thus has the potential to boost access to credit for some of the currently 2 billion working-age adults worldwide who lack access to services in the formal financial sector, thereby fostering financial inclusion and lowering inequality.