NBER WORKING PAPER SERIES

ON THE RISE OF FINTECHS – CREDIT SCORING USING DIGITAL FOOTPRINTS

Tobias Berg
Valentin Burg
Ana Gombović
Manju Puri

Working Paper 24551
http://www.nber.org/papers/w24551

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2018

We wish to thank Frank Ecker, Falko Fecht, Christine Laudenbach, Laurence van Lent, Kelly Shue (discussant), Sascha Steffen, as well as participants of the 2018 RFS FinTech Conference, the 2018 Swiss Winter Conference on Financial Intermediation, and research seminars at Duke University, FDIC, and Frankfurt School of Finance & Management for valuable comments and suggestions. This work was supported by a grant from FIRM (Frankfurt Institute for Risk Management and Regulation). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

On the Rise of FinTechs – Credit Scoring using Digital Footprints
Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri
NBER Working Paper No. 24551
April 2018
JEL No. D12,G20,O33

ABSTRACT

We analyze the information content of the digital footprint – information that people leave online simply by accessing or registering on a website – for predicting consumer default. Using more than 250,000 observations, we show that even simple, easily accessible variables from the digital footprint equal or exceed the information content of credit bureau (FICO) scores. Furthermore, the discriminatory power for unscorable customers is very similar to that of scorable customers. Our results have potentially wide implications for financial intermediaries' business models, for access to credit for the unbanked, and for the behavior of consumers, firms, and regulators in the digital sphere.

Tobias Berg
Frankfurt School of Finance & Management
t.berg@fs.de

Ana Gombović
Frankfurt School of Finance & Management,
a.gombovic@fs.de

Valentin Burg
Humboldt University Berlin
valentin.burg@gmail.com

Manju Puri
Fuqua School of Business
Duke University
100 Fuqua Drive
Durham, NC 27708-0120
and NBER
mpuri@duke.edu

1. Introduction

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". Even without writing text about oneself, uploading financial information, or providing friendship or social network data, the simple act of accessing or registering on a webpage leaves valuable information. As a simple example, every website can effortlessly track whether a customer is using an iOS or an Android device; or track whether a customer comes to the website via a search engine or a click on a paid ad. In this project, we seek to understand whether the digital footprint helps augment information traditionally considered to be important for default prediction and whether it can be used for the prediction of consumer payment behavior and defaults.

Understanding the importance of digital footprints for consumer lending is of significant importance. 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.[1] 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.[2]

In this paper, we analyze the importance of simple, easily accessible digital footprint variables for default prediction using a comprehensive and unique data set covering approximately 250,000 observations from an E-Commerce company located in Germany. Judging the creditworthiness of its customers is important because goods are shipped first and paid later. The use of digital footprints in similar settings is growing around the world.[3] Our data set contains a set of ten digital footprint variables: the device type (for

[1] See in particular Diamond (1984), Boot (1999), and Boot and Thakor (2000) for an overview of the role of banks in overcoming information asymmetries and Berger, Miller, Petersen, Rajan, and Stein (2005) for empirical evidence.

[2] The digital footprint can also be used by financial intermediaries themselves, but to the extent that it proxies for current relationship-specific information it reduces the gap between traditional banks and those firms more prone to technology innovation.

[3] In China, Alibaba's Sesame Credit uses social credit scores from AntFinancial and goods are also shipped first and paid later (see https://www.economist.com/news/finance-and-economics/21710292-chinas-consumer-credit-rating-culture-evolving-fastand-unconventionally-just). Other FinTechs that have publicly announced using digital footprints for lending decisions include ZestFinance and Earnest in the U.S., Kreditech in various emerging markets, and Rapid Finance, CreditEase, and Yongqianbao in China (see https://www.nytimes.com/2015/01/19/technology/banking-start-ups-adopt-new-tools-for-lending.html and https://www.forbes.com/sites/rebeccafeng/2017/07/25/chinese-fintechs-use-big-data-to-give-credit-scores-to-the-unscorable/#45b0e6ed410a).

example, tablet or mobile), the operating system (for example, iOS or Android), the channel through which a customer comes to the website (for example, search engine or price comparison site), a do not track dummy equal to one if a customer uses settings that do not allow tracking device, operating system and channel information, the time of day of the purchase (for example, morning, afternoon, evening, or night), the email service provider (for example, gmail or yahoo), two pieces of information about the email address chosen by the user (includes first and/or last name and includes a number), a lower case dummy if a user consistently uses lower case when writing, and a dummy for a typing error when entering the email address. In addition to these digital footprint variables, our data set also contains data from a private credit bureau that compiles a score similar to the FICO score in the U.S. We are therefore able to assess the discriminatory ability of the digital footprint variables both separately, vis-à-vis the FICO score, and jointly with the FICO score.

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. For example, the difference in default rates between customers using iOS (Apple) and Android (for example, Samsung) is equivalent to the difference in default rates between a median FICO score and the 80^{th} percentile of the FICO score. Bertrand and Kamenica (2017) document that owning an iOS device is one of the best predictors for being in the top quartile of the income distribution. Our results are therefore consistent with the device type being an easily accessible proxy for otherwise hard to collect income data.

Variables that proxy for character and reputation are also significantly related to future payment behavior. For example, customers coming from a price comparison website are almost half as likely to default as customers being directed to the website by search engine ads, consistent with marketing research documenting the importance of personality traits for impulse shopping.[4] Belenzon, Chatterji, and Daley (2017) and Guzman and Stern (2016) have documented an eponymous-entrepreneurs-effect, implying that whether a firm is named after their founders matters for subsequent performance. Consistent with their results, customers having their names in the email address are 30% less likely to default.

[4] See for example Rook (1987), Wells, Parboteeah, and Valacich (2011), and Turkyilmaz, Erdem, and Uslu (2015).

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). The AUC is a simple and widely used metric for judging the discriminatory power of credit scores (see for example Stein, 2007; Altman, Sabato, and Wilson, 2010; Iyer, Khwaja, Luttmer, and Shue, 2016; Vallee and Zeng, 2018). The AUC ranges from 50% (purely random prediction) to 100% (perfect prediction) and is closely related to the Gini coefficient (Gini= 2*AUC–1). The AUC corresponds to the probability of correctly identifying the good case if faced with one random good and one random bad case (Hanley and McNeil, 1982). Following Iyer, Khwaja, Luttmer, and Shue (2016), an AUC of 60% is generally considered desirable in information-scarce environments, while AUCs of 70% or greater are the goal in information-rich environments.

The AUC using the FICO score alone is 68.3% in our data set, comparable to the 66.6% AUC using the FICO score alone documented in a consumer loan sample of a large German bank (Berg, Puri, and Rocholl, 2017), as well as the 66.5% AUC using the FICO score alone in a loan sample of 296 German savings banks (Puri, Rocholl, and Steffen, 2017). As a comparison, Iyer, Khwaja, Luttmer, and Shue (2016) report an AUC of 62.5% in a U.S. peer-to-peer lending data set using the FICO score only. Similarly, in an own analysis we find an AUC of 59.8% using U.S. FICO scores from Lending Club. This suggests that the FICO score provided to us by a German credit bureau clearly possesses discriminatory power and we use the FICO related AUC of 68.3% as a benchmark for the digital footprint variables in our analysis.[5]

Interestingly, a model that uses only the digital footprint variables equals or exceeds the information content of the FICO score: the AUC of the model using digital footprint variables is 69.6%, higher than the AUC of the model using only the FICO score (68.3%). This is remarkable because our data set only contains digital footprint variables that are easily accessible for any firm conducting business in the digital sphere. Our results are also robust to a large set of robustness tests. In particular, we show that digital footprint variables are not simply proxies for time or region fixed effects and results are robust to

[5] Note that the German credit bureau may use some information which U.S. bureaus are legally prohibited to use under the Equal Credit Opportunity Act. Examples include gender, age, current and previous addresses.

various default definitions and sample splits. We also provide out-of-sample tests for all of our results which yield very similar magnitudes. Furthermore, we show that digital footprints today can forecast future changes in the FICO score. This provides indirect evidence that the predictive power of digital footprints is not limited to short-term loans originated online, but that digital footprints matter for predicting creditworthiness for more traditional loan products as well.

In the next step, we analyze whether the digital footprint complements or substitutes for information from the credit bureau. We find that the digital footprint complements rather than substitutes for credit bureau information. The correlation between a score based on the digital footprint variables and the FICO score is only approximately 10%. As a consequence, the discriminatory power of a model using both the FICO score and the digital footprint variables significantly exceeds the discriminatory power of models that only use the FICO score or only use the digital footprint variables. This suggests that a lender that uses information from both sources (FICO + digital footprint) can make superior lending decisions. The AUC of the combined model (FICO + digital footprint) is 73.6% and therefore 5.3 percentage points higher than that of a model using only the FICO score. This improvement is very similar to the 5.7 percentage points AUC improvement reported in Iyer, Khwaja, Luttmer, and Shue (2016) who compare the AUC using the FICO score to the AUC in a setting where, in addition to the FICO score, lenders have access to a large set of borrower financial information as well as access to non-standard information (characteristics of the listing text, group and friend endorsements as well as borrower choice variables such as listing duration and listing category). It is also sizeable relative to the improvement in the AUC by +8.8 percentage points in a consumer loan sample of a large German bank (Berg, Puri, and Rocholl, 2017) and the improvement in the AUC by +11.9 percentage points in a loan sample of 296 German savings banks (Puri, Rocholl, and Steffen, 2017), where the AUC using the FICO score is compared to the AUC using the entire bank-internal information set, including account data, credit history, as well as socio-demographic data and income information. Taken together, this evidence suggests that a few variables from the digital footprint can (partially) substitute for variables that are otherwise more expensive to collect, otherwise take

significantly more effort to provide and process, or might only be available to a few lenders with specific access to particular types of information.

Furthermore, digital footprints can facilitate access to credit when credit bureau scores do not exist, thereby fostering financial inclusion and lowering inequality (Japelli and Pagano, 1993; Djankov, McLiesh, and Shleifer, 2007; Beck, Demirguc-Kunt, and Honohan, 2009; and Brown, Jappelli and Pagano, 2009). We therefore analyze customers for whom no FICO score is available, i.e., customers whose credit history is insufficient to calculate a FICO score, which we label "unscorable customers". We find that the discriminatory power of the digital footprint for unscorable customers matches the discriminatory power for scorable customers (72.2% versus 69.6% in-sample, 68.8% versus 68.3% out-of-sample). These results suggest that digital footprints have the potential to boost financial inclusion to parts of the currently two billion working-age adults worldwide that lack access to services in the formal financial sector.

In the last section, we discuss implications of our findings for the behavior of consumers, firms and regulators. Consumers might plausibly change their behavior if digital footprints are widely used for lending decisions (Lucas (1976)). Some of the digital footprint variables are clearly costly to manipulate (such as buying the newest smart device or signing up for a paid email account) while others require a customer to change her intrinsic habits (such as impulse shopping or making typing mistakes). However, more importantly, such a change in behavior can lead to a situation where consumers fear to express their individual personality online. A wider implication of our findings is therefore that the use of digital footprints has a considerable impact on everyday life, with consumers constantly considering their digital footprints which are so far usually left without any further thought. Firms and regulators are equally likely to react to an increased use of digital footprints. As an example, firms associated with low creditworthiness products may object to the use of digital footprints and may conceal the digital footprint of their products. Regulators are likely to watch closely whether digital footprints proxy for variables that are legally prohibited to be used for credit scoring.

Our paper relates to the literature on the role of financial intermediaries in mitigating information asymmetries (Diamond, 1984; Petersen and Rajan, 1994, Boot, 1999; Boot and Thakor, 2000; Berger,

Miller, Petersen, Rajan, and Stein, 2005). The prior literature has established the importance of credit history and account data to assess borrower risk (Mester, Nakamura, and Renault, 2007; Norden and Weber, 2010; Puri, Rocholl, and Steffen, 2017), thereby giving rise to an informational advantage for those financial intermediaries with access to borrowers' credit history and account data. More recently, the literature has explored the usefulness of data beyond the FICO score and bank-internal relationship-specific data for default prediction. These data sources include soft information in peer-to-peer lending (Iyer, Khwaja, Luttmer, and Shue, 2016), friendships and social networks (Hildebrandt, Rocholl, and Puri, 2017; Lin, Prabhala, and Viswanathan, 2013), text-based analysis of applicants listings (Gao, Lin, and Sias, 2017; Dorfleitner et al., 2016), and signaling and screening via contract terms (reserve interest rates in Kawai, Onishi, and Uetake 2016; maturity choice in Hertzberg, Liberman, and Paravisini, 2017).

Our paper differs from these papers, in that the information we are looking at is provided simply by accessing or registering on the website, not by furnishing any information – hard or soft – about the applicant. We show that even simple, easily accessible variables from the digital footprint provide valuable information for default prediction that helps to significantly improve traditional credit scores. Our variables stand out in terms of their ease of collection: almost every firm operating in the digital sphere can effortlessly track the digital footprint we use. Unlike the papers cited above, the processing and interpretation of these variables does not require human ingenuity, nor does it require effort on the side of the applicant (such as uploading financial information or inputting a text description about oneself), nor does it require the availability of friendship or social network data. Simply accessing or registering on the website is adequate. Our results imply that barriers to entry in financial intermediation might be lower in a digital world, and easily accessible digital footprints can (partially) substitute for variables that need to be collected with considerable effort in a non-digital world. As a consequence, the digital footprint can also be used to process applications faster than traditional lenders (see Fuster et al. (2018) for an analysis of process time of FinTech lenders versus 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 that might either be more costly to collect or only accessible to a selected group of intermediaries.

The rest of the paper is structured as follows. Section 2 provides an overview about the institutional setup and data. Section 3 provides empirical results. Section 4 discusses further implications of our findings. Section 5 concludes.

2. Institutional setup, descriptive statistics, and the digital footprint

2.1 Institutional setup

We access data about 270,399 purchases from an E-commerce company selling furniture in Germany (similar to "Wayfair" in the U.S.) between October 2015 and December 2016. Before purchasing an item, a customer needs to register using his or her name, address and email. Judging the creditworthiness of its customers is important because goods are shipped first and paid later.[6] The claims in our data set are therefore akin to a short-term consumer loan.

The company uses information from two private credit bureaus to decide whether customers have a sufficient creditworthiness. The first credit bureau provides basic information such as whether the customer exists and whether the customer is currently or has been recently in bankruptcy. This score is used to screen out customers with fraudulent data as well as customers with clearly negative information.[7] The second credit bureau score draws upon credit history data from various banks (credit card debt and loans outstanding, past payment behavior, number of bank accounts and credit cards), sociodemographic data, as well as payment behavior data sourced from retail sales firms, telecommunication companies, and utilities. This second credit bureau score is similar to the FICO score in the U.S. and we will label this score "FICO score" for ease of understanding. This FICO score is requested for purchases exceeding € 100 and we consequently restrict our data set to purchases for which the company requested a FICO score.[8] We label those customers for whom a FICO exists "scorable customers".

[6] Customers can choose to pay upfront instead of paying after shipment of the products. Customers paying upfront are not included in our data set. Paying after shipment, so called "deferred payment", is by far the dominant payment type: more than 80% of customers choose to pay after shipment if this method is offered to a customer.

[7] The firm switched the credit bureau that provides this basic information in July 2016. Results are very similar pre-July-2016 and post-July-2016.

[8] The company requests the FICO score if the customer's shopping cart amount exceeds € 100, even when the customer ultimately purchases a smaller amount.

The E-commerce company uses the FICO score together with the digital footprint (discussed further below) to screen out borrowers with a predicted default rate exceeding 10 percent. Restricting our data set to orders exceeding € 100 and excluding customers with a very low creditworthiness has the benefit of making our data set more comparable to a typical credit card, bank loan or peer-to-peer lending data set.

After the purchase, the items are sent to the customer together with an invoice. The customer has 14 days to pay the invoice. If the customer does not pay on time, three reminders (one per email, two per email and letter) are sent out. 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.

2.2 Descriptive statistics

Our data set comprises 270,399 purchases between October 2015 and December 2016. The FICO score is available for 254,808 observations (94% of the sample) and unavailable for 15,591 observations (6% of the sample). Non-existence is due to customers being unscorable, i.e., not having a sufficient credit history that would allow the credit bureau to calculate a FICO score. In the following and throughout the entire paper, we distinguish between scorable and unscorable borrowers, i.e. those with and without FICO score. As shown in Figure 1a, the purchases are distributed roughly even over time with slight increases in orders during October and November, as typical for the dark season. Table 2 provides descriptive statistics for both subsamples, variable descriptions are in Table 1.

[Table 1 and Table 2, Figure 1a, Figure 1b, Figure 2]

In the sample with FICO score, the average purchase volume is EUR 318 (approximately USD 350) and the mean customer age is 45.06 years. On average, 0.9% of customers default on their payment. Our default definition comprises claims that have been transferred to a debt collection agency.[9] The FICO score ranges from 0 (worst) to 100 (best). It is highly skewed with 99% of the observations ranging

[9] The average time between the order date and the date a claim is transferred to the debt collection agency is 103 days in our sample, i.e., approximately 3.5 months.

between 90 and 100. The average FICO score is 98.11, the median is 98.86. Figure 2 provides the distribution of FICO scores together with (smoothed) default rates. The average FICO score of 98.11 corresponds to a default rate of approximately 1% and default rates grow exponentially when FICO scores decrease, with a FICO of 95 corresponding to a 2% default rate and a FICO of 90 corresponding to a 5% default rate. Note that default rates are not annualized but constitute default rates over a shorter window of approximately 3.5 months.

Descriptive statistics for the sample without FICO 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) and default rates being significantly higher (2.5%).

2.3 Representativeness of data set

Our data set is largely representative of the geographic distribution of the German population overall. As can be seen from Figure 1b, the share of observations in our sample closely follows the population share for all the 16 German states. Furthermore, the mean customer age is 45.06 years, comparable both to the mean age of 43.77 in the German population as well as to the mean age of 45.24 reported by Berg, Puri, and Rocholl (2017) in a sample of more than 200,000 consumer loans at a large German private bank. Our sample is restricted to customers of legal age (18 years and older) and less than 5% of the customers are older than 70. The age distribution in our sample therefore resembles the age distribution of the German population aged 18-70: the interquartile range of the German population aged 18-70 ranges from 31-56, compared to an interquartile range of 34-54 in our sample.

The average default rate in our sample is 1.0% (0.9% for scorable customers, 2.5% for unscorable customers). As discussed above, these default rates constitute default rates over a window of approximately 4 months, implying a scaled-up annualized default rate of 3.0%. We compare our default rate to other studies in Appendix Table A.1. Berg, Puri, and Rocholl (2017) report an average default rate of 2.5% in a sample of more than 200,000 consumer loans at a large German private bank; the major German credit bureau reports an average default rate of 2.4% (2015) and 2.2% (2016) in a sample of more than 17 million

consumer loans, and the two largest German banks report probability of default estimates of 1.5% (Deutsche Bank) and 2.0% (Commerzbank) across their entire retail lending portfolio. Default rates reported by Puri, Rocholl, and Steffen (2017) in a sample of German savings banks are somehow lower. Taken together, this evidence suggests that default rates in our sample are largely representative of a typical consumer loan sample in Germany. Charge-off rates on consumer loans in the U.S. across all commercial banks as reported by the Federal Reserve were approximately 2% in 2015/2016, implying a comparable default rate to our sample. Default rates reported in some U.S. peer-to-peer lending studies are higher (up to 10% per annum). However, the studies with the highest default rates were conducted using loans originated in 2007/2008 at the height of the financial crisis. More recent studies report default rates that are comparable to our default rates on an annualized basis (for example, Hertzberg, Liberman, and Paravisini, (2016) report a 4.2% annualized default rate in a sample of Lending Club loans originated in 2012/2013).

2.4 Digital footprint

In addition to the credit bureau score described above, the company collects a "digital footprint" for each customer. All digital footprint variables are simple, easily accessible variables that every firm operating in the digital sphere can collect at almost no cost. The list of all digital footprint variables is reported in Table 1.

The digital footprint comprises easily accessible pieces of information known to be a proxy for the economic status of a person, for instance the device type (desktop, tablet, mobile) and operating system (for example, Windows, iOS, Android). As documented by Bertrand and Kamenica (2017), owning an iOS device is one of the best predictors for being in the top quartile of the income distribution. Furthermore, the distinct features of most commonly used email providers in Germany (for example Gmx, Web, T-Online, Gmail, Yahoo, or Hotmail) also allow us to infer information about the customer's economic status. Gmx, Web, and T-online are common email hosts in Germany which are partly or fully paid. In particular, T-online is a large internet service provider and is known to serve a more affluent clientele, given that it offers internet, telephone, and television plans and in-person customer support. A customer obtains a T-

online email address only if she purchased a T-online package. Yahoo and Hotmail, in contrast, are fully free and mostly outdated services. Thus, based on these simple variables, the digital footprint provides easily accessible proxies of a person's economic status absent of private information and hard-to-collect income data.

Second, the digital footprint provides simple variables known to proxy for character, such as the channel through which the customer has visited the homepage of the firm. Examples for the channel include paid clicks (mainly through paid ads on google or by being retargeted by ads on other websites according to preferences revealed by prior searches), direct (a customer directly entering the URL of the E-commerce company in her browser), affiliate (customers coming from an affiliate site that links to the E-commerce company's webpage such as a price comparison site), and organic (a customer coming via the non-paid results list of a search engine). Information about a person's character (such as her self-control) is also reasonably assumed to be revealed by the time of day at which the customer makes the purchase (for instance, we find that customers purchasing between noon and 6 pm are approximately half as likely do default as customers purchasing from midnight to 6am).

Finally, corporate research documents that firms being named after their owners have a superior performance. This so called eponymous effect is mainly driven via a reputation channel (Belenzon, Chatterji, and Daley, 2017). We find it reasonable to extend this finding to the choice of email addresses. A testable prediction from this prior literature is that eponymous customers – those who include their first and/or last names in their email address – are less likely to default. In contrast to eponymous customers, those arguably less concerned with including their name but instead include numbers or type errors in their email address default more frequently.[10] The digital footprint provides this type of simple information that can serve as a proxy for reputation in the form of four dummies, as to whether the last and/or first name is part of the email address, whether the email address contains a number, whether the email contains an

[10] Approximately 10-15% of defaults are identified as fraud cases. Compared to non-fraud defaults, fraud cases have a higher incidence of numbers in their email address. This is consistent with anecdotal evidence suggesting that fraudsters create a large number of email addresses and do so in a way that uses a string combined with consecutive numbers.

error, as well as whether the customer types either the name or shipping address using lower case on the homepage.[11]

Note that some of the variables discussed above are likely to proxy for several characteristics. For example, iOS devices are a predictor of economic status (Bertrand and Kamenica, 2017), but might also proxy for character (for example, status-seeking users might be more likely to buy an iOS device). It is not our target to point to exactly one single channel that can explain why digital footprints variables can predict default. Rather we want to highlight existing research that provides guidance as to why we can expect these variables to matter for default prediction.

3. Empirical results

3.1 Univariate results

We provide univariate results for the sample of customers with FICO score in Table 3.

[Table 3]

As expected, the FICO score clearly exhibits discriminatory ability: the default rate in the lowest FICO quintile is 2.12%, more than twice the average default rate of 0.94% and five times the default rate in the highest FICO quintile (0.39%).[12]

Interestingly, the univariate results indicate discriminatory ability for the digital footprint variables as well. The footprint variables that proxy for income and wealth reveal significant differences in payment behavior. For example, orders from mobile phones (default rate 2.14%) are three times as likely to default as orders from desktops (default rate 0.74%) and two-and-a-half times as likely to default as orders from

[11] Kreditech is an example of a German company already using simple typography variables, such as the lack of capital letters, to evaluate credit risk but also detect possible fraud and online impersonations (see BBVA (2017): The digital footprint: a tool to increase and improve lending, accessed via https://www.bbva.com/en/digital-footprint-tool-increase-improve-lending/).

[12] Using U.S. FICO scores from Lending Club over the same period we find that the default rate increases only by a factor of 2.5 from the highest to the lowest FICO quintile, suggesting our FICO score has more discriminatory power than the U.S. FICO, which we will confirm later using AUCs.

tablets (default rate 0.91%). Orders from the Android operating systems (default rate 1.79%) are almost twice as likely to default as orders from iOS systems (1.07%) – consistent with the idea that consumers purchasing an iPhone are usually more affluent than consumers purchasing other smartphones. As expected, customers from a premium internet service (T-online, a service that mainly sells to affluent customers at higher prices but with better service) are significantly less likely to default (0.51% versus the unconditional average of 0.94%). Customers from shrinking platforms like Hotmail (an old Microsoft service) and Yahoo exhibit default rates of 1.45% and 1.96%, almost twice the unconditional average.

Information on character is also significantly related to default rates. Customers arriving on the homepage through paid ads (either clicking on paid google ads or being retargeted after prior google searches) exhibit the largest default rate (1.11%). One possible interpretation is that ads, in particular ads that are shown multiple times on various websites to a customer, seduce customers to buy products they potentially cannot afford. Customers being targeted via affiliate links, e.g. price comparison sites, and customers directly entering the URL of the E-commerce company in their browser exhibit lower-than-average default rates (0.64% and 0.84%). Finally, customers ordering during the night have a default rate of 1.97%, approximately two times the unconditional average.

There are only few customers who make typing mistakes while inputting their email addresses (roughly 1% of all orders), but these customers are much more likely to default (5.09% versus the unconditional mean of 0.94%). Customers with numbers in their email-addresses default more frequently, which is plausible given that fraud cases also have a higher incidence of numbers in their email address.[13] Customers who use only lower case when typing their name and shipping address are more than twice as likely to default as those writing names and addresses with first capital letters. Interestingly, we find that eponymous customers who use their first and/or last name in their email address are less likely to default. Thus information on reputation also shows significant power for predicting default rates. These findings are

[13] Approximately 10-15% of defaults are identified as fraud cases. Compared to non-fraud defaults, fraud cases have a higher incidence of numbers in their email address. This is consistent with anecdotal evidence suggesting that fraudsters create a large number of email addresses and do so in a way that uses a string combined with consecutive numbers.

consistent with recent findings by Belenzon, Chatterji, and Daley (2017) who show that eponymous firms perform better, supporting the reputational explanation of their findings.

3.2 Measures of association between variables, Combination of digital footprint variables

In the next step, we report measures of association between the FICO score and the digital footprint variables in order to assess whether the digital footprint variables are correlated with the FICO score and among each other, or whether they provide independent information. As most of the digital footprint variables are categorical variables, standard measures for ordinal variables (for example, Pearson's correlation or Spearman rank correlation) are not feasible. We therefore report Cramér's V, which provides a measure of association between categorical variables that is bounded in the interval [0,1], with 0 denoting no association and 1 denoting perfect association. To allow calculation of Cramér's V, we transform the continuous variables (FICO score and Check-Out Time) into categories by forming quintiles by FICO score and categorizing the check-out time into morning, afternoon, evening, and night. Table 4 reports the results.

[Table 4]

Interestingly, the Cramér's V between the FICO 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 FICO scores – a claim we will analyze more formally below in a multivariate regression setup.

The association between the variables Device Type and Operating System is high. This is not surprising, for example most desktop computers run on Windows and most tablets on iOS or Android. To avoid multicollinearity, we therefore simply use the most frequent combinations from these two categories in our multivariate regressions below.[14] The check-out time has some association with device type/operating system. Mobile phones are used relatively more frequently than desktops and tablets for late

[14] The most frequent combinations are Windows and Macintosh for desktop computers, Android and iOS for tablets, and Android and iOS for mobile phones. See Table A.3 in the Appendix for descriptive statistics.

night shopping, and desktops are used relatively more in the afternoon. All other combinations of digital footprint variables have a Cramér's V of less than 0.25.

The fact that many of the digital footprint variables provide mutually independent information suggests that a combination of digital footprint variables is significantly more powerful in predicting default than single variables. We illustrate this idea in Figure 3. Figure 3 depicts default rates using the variables "Operating system" and "Email host" separately as well as in combination. The sample is restricted to customers with FICO score.

[Figure 3]

Among the categories from these two variables, T-online users have the lowest default rate (0.51%), while Yahoo users have the highest default rate (1.96%). As a reference point, we list deciles by FICO score at the bottom of Figure 3. The default rate of T-online users of 0.51% is approximately equal to the default rate in the 7th decile of FICO scores, while the default rate of Yahoo users (1.96%) is between the 1st and 2nd decile of FICO scores. When combining information from both variables ("Operating system" and "Email host"), default rates are even more dispersed.[15] We observe the lowest default rate for Mac-users with a T-online email address. The default rate for this combination is 0.36%, which is lower than the average default rate in the 1st decile of FICO scores. On the other extreme, Android users with a Yahoo email address have an average default rate of 4.30%, significantly higher than the 2.69% default rate in the highest decile of FICO scores. These results suggest that even two simple variables from the digital footprint allow categorizing customers into default bins that match or exceed the variation in default rates from FICO deciles.

3.3 Multivariate results: Digital footprint and default

Table 5 provides multivariate regression results of a default dummy on the FICO score and digital footprint variables. We use a logistic regression and report the Area-Under-Curve (AUC) for every

[15] The following results are not driven by small sample sizes, i.e., all categories reported in Figure 3 have at least 1,000 observations.

specification. The AUC is a simple and widely used metric for judging the discriminatory power of credit scores (see for example Stein, 2007; Altman, Sabato, and Wilson, 2010; Iyer, Khwaja, Luttmer, and Shue, 2016). The AUC ranges from 50% (purely random prediction) to 100% (perfect prediction). Following Iyer, Khwaja, Luttmer, and Shue (2016), an AUC of 60% is generally considered desirable in information-scarce environments, while AUCs of 70% or greater are the goal in information-rich environments. We also plot the Receiver Operating Characteristic that is used to calculate the AUC in Figure 4.

[Table 5 and Figure 4]

Column (1) of Table 5 reports results using the (continuous) FICO score as an independent variable. As expected and consistent with Figure 2, the FICO score is a highly significant predictor of default, with higher FICO scores being associated with lower default rates. The AUC using only the FICO score is 68.3% and is significantly different from chance (AUC of 50%). This result is comparable to the 66.6% AUC using the FICO score alone documented in a consumer loan sample of a large German bank (Berg, Puri, and Rocholl, 2017) and the 66.5% AUC using the FICO score alone in a loan sample of 296 German savings banks (Puri, Rocholl, and Steffen, 2017). This result is higher than the AUC of 62.5% reported by Iyer, Khwaja, Luttmer, and Shue (2016) in a U.S. peer-to-peer lending data set using the FICO score only and the AUC of 59.8% we compute for comparison using U.S. FICO scores from Lending Club. This suggests that the FICO score provided to us by a German credit bureau clearly possesses discriminatory power and we use the AUC of 68.3% as a benchmark for the digital footprint variables in the following.

Column (2) reports results for the digital footprint, column (3) uses both the FICO score and the digital footprint variables, and column (4) adds age and month and region fixed effects. 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. We report AUCs in the bottom rows of Table 5 and also test for differences in AUCs using the methodology by DeLong, DeLong, and Clarke-Pearson (1988).

Interestingly, digital footprint variables have an AUC of 69.6% – which is higher than the AUC of the FICO score.[16] These results suggest that even simple, easily accessible variables from the digital footprint are as useful in predicting defaults as the FICO score. We focus on the economic and statistical significance of the variables in column (2) in the following discussion.

The variables "Email error", "Mobile/Android", and the "Night" dummy have the highest economic significance. The variable "Email error" is a simple dummy variable that is equal to one in only a few cases, and thus allows categorizing a small portion of customers as being high risk. Customers with an Email Error have an odds ratio of defaulting which is exp(1.66)=5.25 times higher than customers without an Email Error. Given that default rates are rather small, default probabilities *p* and odds ratios (*p/(1-p)*) are very similar, implying that customers with an Email Error default approximately 5.25 times more frequent than customers without Email Error.

Android users default more frequently than the baseline category, consistent with the univariate results and consistent with the fact that consumer purchasing an iPhones are usually more affluent than consumers purchasing other smartphones. Customers purchasing at night (midnight-6am) also default more frequently than customers purchasing at other times of the day, suggesting that purchases made during a time when many people might be asleep are fundamentally different from daytime purchases.

In column (3) of Table 5, we complement the digital footprint variables with the FICO score. Both the coefficient on the FICO score as well as the coefficients on the digital footprint variables barely change compared to columns (1) and (2). This suggests that the digital footprint variables complement rather than substitute for the information content of the FICO score. As a consequence, the AUC of the combined model using both the digital footprint variables and the FICO score (73.6%) is significantly higher than the AUC of each of the stand-alone models (68.3% for the FICO score and 69.6% for the digital footprint variables).[17]

[16] Note that in Table 5 we report only the 6 largest categories for email providers even though we use the largest 18 categories in the regression (all email providers with at least 1000 observations). In a regression using only the 6 reported email hosts, the AUC of the digital footprint decreases by 0.9PP, still being higher than the AUC using FICO alone.

[17] Please note that AUCs generated by two independent variables cannot be simply summed up because the AUC of an uninformative variable is already 50%.

In column (4) of Table 5, we add time and region fixed effects and control for age. Results remain almost unchanged, suggesting that neither the FICO score nor the digital footprint act as simple proxies for different regions, different sub-periods, or different age. While older people are expectedly less likely to default, consistent with the idea that it takes time to build up a credit history, coefficients for the FICO score and the digital footprint remain very similar.[18] Only the coefficient for users of the premium service T-online, which is known to serve more affluent and older customers, decreases slightly in economic significance (from -0.35 in column (3) to -0.27 in column (4)).

Figure 5 provides a more detailed look at the correlation between the FICO score and the digital footprint. Using the results from column (2) of Table 5, we construct a default prediction using only the digital footprint variables for each observation in our sample. For each observation, Figure 5 then depicts the percentile using the FICO score as well as the percentile using the digital footprint score. As an example, if a customer has a very good FICO score (=low default probability) and a very low default probability by the digital footprint as well, then it would end up in the upper right-hand corner of Figure 5. A customer with a low FICO score (=high default probability) and a very high default probability by the digital footprint as well would end up in the lower left-hand corner. Observations where FICO score and digital footprint have opposing predictions end up in the upper left-hand corner or the lower right-hand corner. Figure 4 clearly shows that the correlation between FICO score and digital footprint is very low (R^2 of 1.3%, 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 FICO score.

[Figure 5]

3.4 Unscorable customers and access to finance

[18] The coefficient on log(age) in Column (4) of Table (5) is -0.33 (significant at the 1% level), suggesting that a doubling in age reduces defaults by approximately one third.

The lack of access to financial services affects around two billion working-age adults worldwide and is seen as one of the main drivers of inequality.[19] Particularly in developing countries, the inability of the unbanked population to participate in financial services is often caused by a lack of information infrastructure, such as credit bureau scores. Many countries have therefore already started leveraging digital technologies to promote financial inclusion.[20] As digital footprint variables are available also for customers without a credit score, analyzing borrowers' digital behavior may present an opportunity to boost financial inclusion – particularly in developing economies, where a large share of the population does not have access to banking services and therefore to traditional credit sources.[21]

We test whether the digital footprint can present an opportunity to facilitate access to finance for customers who do not have a credit bureau score, which we label unscorable customers in our analysis. The average default rate of unscorable customers in our sample is 2.49% (see Table 6), thereby clearly exceeding the default rate for scorable customers of 0.94% (see Table 2). This is not surprising, given that unscorable customers are customers without credit record where uncertainty about repayment is likely to be higher. Interestingly, the discriminatory power of the digital footprint – as measured by the AUC – is broadly similar for unscorable customers than for scorable customers (72.2% versus 69.6%), see Table 7 and Figure 6. Adding time and region fixed effects also does not affect our results (Column (3) of Table 7).

[Table 6, Table 7 and Figure 6]

These results suggest that digital footprints may help to overcome information asymmetries between lenders and borrowers when standard credit bureau information is not available. We clearly have to be cautious in interpolating these results from a developed country to unscorable customers in emerging markets. Still, recent activity in the FinTech industry suggests this is an avenue that FinTechs aim to take.

[19] The World Bank Group identifies financial inclusion as a key enabler of reducing poverty and boosting prosperity and promotes new use of data and digital technology as an opportunity for expanding access to financial services. See e.g. http://www.worldbank.org/en/news/video/2016/03/10/2-billion-number-of-adults-worldwide-without-access-to-formal-financial-services, http://www.worldbank.org/en/topic/financialinclusion

[20] Initiatives include the G20 High-Level Principles for Digital Financial Inclusion available via https://www.gpfi.org/sites/default/files/G20%20High%20Level%20Principles%20for%20Digital%20Financial%20Inclusion.pdf

[21] See BBVA (2017), suggesting analyzing borrowers' online behavior can help financial inclusion particularly in emerging economies, available via https://www.bbva.com/en/digital-footprint-tool-increase-improve-lending/

Motivated by a dramatic increase in the availability of digital footprints in developing economies, new FinTech players have emerged that use digital footprints to challenge traditional banking options and develop innovative financing solutions.[22] These FinTechs 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 (see Japelli and Pagano, 1993; Djankov, McLiesh, and Shleifer, 2007; Beck, Demirguc-Kunt, and Honohan, 2009; and Brown, Jappelli and Pagano, 2009 for the link between availability of credit scores, access to credit and inequality).

3.5 Out-of-sample tests

Table 5 (scorable customers) and Table 6 (unscorable customers) were estimated in-sample which may overstate discriminatory power due to overfitting. We therefore provide out-of-sample tests using Nx2-fold cross validation in Table 8. Nx2-fold cross validation is a common method to evaluate out-of-sample performance of an estimator (see for example Dietterich, 1998 for a general discussion of cross-validation techniques). We thereby randomly divide the full sample into half samples A and B, estimate a predictive logistic regression using sample A, and use the coefficients to create predicted values for the observations in sample B. We then estimate a predictive regression using sample B and use the coefficients to create predicted values for observations in sample A. Finally, we determine the AUC for the full sample of observations, using all predicted values estimated out-of-sample. We repeat this procedure N=100 times and report the mean out-of-sample AUCs in column (2) of Table 8.

Panel 1 of Table 8 reports out-of-sample results for scorable customers. The out-of-sample AUC is less than 1 PP lower than the in-sample AUC for all specifications apart from the fixed effects regression. In the fixed effects specification, where we insert granular month and region fixed effects, out-of-sample AUCs are 2.8 PP lower than in-sample AUCs. This is not surprising given that overfitting is in particular an issue when many explanatory variables are used. AUCs for the fixed effects regressions are of little

[22] See e.g. https://hbr.org/2017/01/fintech-companies-could-give-billions-of-people-more-banking-options

relevance for our paper as the fixed effects regressions serve the sole purpose of showing that neither the FICO score nor the digital footprint variables are simple proxies for month or region fixed effects.

Panel 2 of Table 8 reports out-of-sample results for unscorable customers. The out-of-sample AUC using Nx2-fold cross validation decreases by 3.9 PP for unscorable customers compared to the in-sample estimate. The sample size for unscorable customers is significantly smaller than for scorable customers so that a larger gap between out-of-sample and in-sample AUCs is expected. The out-of-sample AUC of the digital footprint for unscorable customers (68.3%) is, however, still similar to the out-of-sample AUC of 68.8% for scorable customers. Overall, our main conclusion – digital footprints provide a similar predictive power for unscorable customers than for scorable customers – is clearly confirmed in out-of-sample tests as well.

[Table 8]

3.6 Alternative default definitions and sample splits

Table 9 provides various robustness tests. Panel A uses alternative default definitions and Panel B provides results for various sample splits. In all Panels, we report the area under curve (AUC) for the FICO score, for the digital footprint, and for both together.

Panel A uses an alternative default definition, namely default after efforts by the collection agency, in column (2). The collection agency is able to fully recover approximately 40% of the claims, resulting in a reduced default rate after the collection agency process. The relative importance of FICO versus digital footprint is almost unaffected and the AUC increases slightly. This seems intuitive, given that it is harder to predict customers who don't pay in the first months, but pay at a later point in time than to simply predict customers that won't be able to pay at all. Column (3) of Panel A reports results using the loss given default as the dependent variable. Compared to the FICO score, the digital footprint is both economically and statistically a better predictor of loss given default. The digital footprint therefore does not only help to predict default, but also helps to predict recovery rates for defaulted exposures. Panel B reports various

sub-sample splits. Results are very similar for small and large orders (split at the median) as well as for female and male customers.

Overall, the robustness tests suggest that our key results from Table 5 – digital footprints predict default as well or even better than the FICO score, and digital footprint and FICO score being complements rather than substitutes, is robust for different default definitions and various sample splits. This suggests even simple, easily accessible variables from the digital footprint are important for default prediction over and above the information content of credit bureau (FICO) scores.

[Table 9]

3.7 External validity

The evidence presented so far provides evidence of the predictive power of the digital footprint for short term loans for products purchased online. In Section 2.3 (and Appendix Table A.2) we have shown that our data set is largely representative to a typical German consumer loan sample in terms of age distribution, geographic distribution, as well as default rates. Appendix Table A.2 – also discussed in Section 2.3 – further shows that the FICO score has a very similar predictive power in our sample compared to consumer loan samples both at German savings banks as well as at German private banks.

In this section, we provide further evidence for the external validity of our setting. In particular, we test whether digital footprints today can forecast future changes in the FICO score. If a good digital footprint today predicts an increase in the FICO 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:

$$\Delta(FICO_{t+1}, FICO_t) = \beta_0 + \beta_1 \Delta(DF_t, FICO_t) + X + \varepsilon \quad (1)$$

where $\Delta(FICO_{t+1}, FICO_t)$ is the change in FICO between t+1 and t, $\Delta(DF_t, FICO_t)$ is the difference between predicted default rates using the digital footprint variables (i.e., predicted values from column (2) of Table 5) and predicted default rates using the FICO score (i.e., predicted values from column (1) of Table 5), and X is a set of control variables. We winsorize both the dependent and the independent variable in equation (1) at the 1/99 percent level. A limitation of our dataset is that the left-hand side variable is

available only for customers that are part of our original dataset and have returned to the E-Commerce company at least once up to March 2018.[23] For each observation in our original data set from Table 5 we check whether the customer returned to the platform and report the latest available FICO score for each customer. For returning customers, the E-Commerce company only requests a new FICO score if the existing FICO score is older than six months, implying that the difference between t and t+1 in equation (1) is at least 181 days. The average (median) time between t and t+1 in equation (1) is 450 days (431 days), i.e. a little over one year.[24]

Figure 7 provides descriptive evidence of the predictive power of the digital footprint for the subsequent development of FICO scores. We split our sample into 10 deciles by FICO score and further split each of the 10 decile into observations where the digital footprint predicts a lower probability of default than the FICO score (grey line in Figure 7) and observations where the digital footprint predicts a higher probability of default than the FICO score (black line in Figure 7). We then plot the average of the subsequent FICO score on the vertical axis for each of these 20 bins (10 deciles x digital footprint better/worse than FICO). Figure 7 shows that the grey line is consistently above the black line, suggesting that customers with better digital footprints also see a better development of their FICO score in the future.

[Figure 7 and Table 10]

Table 10 provides formal regression results. Column (1) provides results without control variables. The coefficient on $\Delta(DF_t, FICO_t)$ is economically and statistically highly significant. The coefficient of -74.56 suggests that if the digital footprint default prediction is 1PP higher than the FICO default prediction

[23] The data set in Table 5 is limited to the period from October 2015 to December 2016 to allow for a subsequent observation of default rates and loss given defaults. For changes in FICO scores we expand the data set until March 2018. Please note that while the sample from Table 5 is limited to customers that pass the minimum-creditworthiness condition (see Section 2.1), the subsequent FICO score is also available for returning customers that were denied buying via invoice upon returning due to a very low FICO score.

[24] It is plausible that changes in FICO scores affect customers' decision to return to the E-Commerce company, but such a selection does not necessarily invalidate our regression design. For the estimate of β_1 in equation (1) we rely on the assumption that the decision to return to the E-Commerce platform is not related to both the difference $\Delta(DF_t,FICO_t)$ *and* the subsequent change in FICO scores. If, for example, customers whose creditworthiness using the digital footprint is better than their creditworthiness using the FICO score return only if their FICO score has increased, then the coefficient β_1 would be downward biased (and vice versa).

(for example, the digital footprint predicts a 2% default probability while the FICO predicts a 1% default probability), then the FICO score decreases by 0.75 points in the future. Given that German FICO scores represent 1-year survival probabilities, this suggests that the FICO score adjusts 75% on its way towards the digital footprint. Figure 7 shows some mean reversion in FICO scores, with low FICO scores getting better on average and high FICO scores getting worse. To ensure that our results are not purely driven by mean-reversion, we control for $FICO_t$ in column (2). As expected, the coefficient decreases but remains both economically and statistically highly significant at -28.14. Controlling for month and region fixed effects barely changes the coefficient (column (3) of Table 10). The effect is rather monotone across quintiles by $\Delta(DF_t, FICO_t)$, suggesting that effects are not driven only by particularly negative or particularly positive digital footprints. Consistent with Figure 7, there is some evidence that the effects are somehow larger for lower FICO scores (see column (5) and (6) of Table 10), but the digital footprint clearly possesses predictive power for future changes in the FICO score for higher FICO scores as well.

Taken together, the evidence suggests that digital footprints today forecast subsequent changes in FICO scores. This result provides a window into the traditional banking world. As FICO 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.

4. Implications for the behavior of consumers, firms, and regulators

Implications for the behavior of consumers

Our prior results are subject to the Lucas (1976) critique, with customers potentially changing their online behavior if digital footprints are widely used in lending decisions. Some of the digital footprint variables are clearly costly to manipulate (such as buying the newest smart device or signing up for a paid email account) while others require a customer to change her intrinsic habits (such as impulse shopping or making typing mistakes). In the following we lay out two major implications of the Lucas critique: one that

directly affects the use of digital footprints in lending decisions, and another implication that affects the right to free development and expression of one's personality more generally.

First, in the long-run, the discriminatory power of the digital footprint depends on how easily bad types can mimic good types. If mimicking good types is costless, an uninformative pooling equilibrium evolves. A sufficiently high cost for mimicking good types results in a separating equilibrium, possibly making the digital footprint even more informative than is currently the case (Spence (1973)). Clearly, other digital footprint variables may evolve in the future that are costly to manipulate or proxy for a person's innate character. A particular vibrant example is Pentaquark's scoring model that rejects loans from applicants who write a lot about their souls on Facebook, as these people are usually too concerned about what will happen in thirty years, but not the fine print of today's life.[25]

Second, and even more importantly, the repercussions of our findings seem even stronger when consumers adapt their behavior in a conformable way. A world of conformity, where consumers fear to express their individual personality and act with a permanent desire to portray a positive image to others is not the role model of a society that most people think of. This argument becomes even more relevant if lenders expand the scope of digital footprint variables they use, the more of our devices are connected to the internet, and the more of our personal communication can be traced online. A wider implication of our findings is therefore that the use of digital footprints has a considerable impact on everyday life, with consumers constantly considering their digital footprints which are so far usually left without any further thought.

Implications for the behavior of firms and regulators

The Lucas critique applies not only to consumer behavior, but firms and regulators are equally likely to react to an increased use of digital footprints. Firms associated with low creditworthiness products may object to the use of digital footprints and may conceal the digital footprint of their products. Commercial services that offer to manage individual's digital footprints may develop. On the other hand,

[25] See BBVA (2017): The digital footprint: a tool to increase and improve lending, accessed via https://www.bbva.com/en/digital-footprint-tool-increase-improve-lending/

firms whose products are pooled with low-quality products (for example high-cost Android phones) may want to ensure that their digital footprint clearly distinguishes them from lower-reputation products in the same category. Overall, similar to our discussion on consumer behavior, the reaction by firms can either increase or decrease the predictive power of the digital footprint.

Regulators are likely to watch the use of digital footprints closely. Regulators worldwide have long recognized the key role of credit scores for consumers' access to key financial products. As a consequence, lending acts worldwide – such as the Equal Credit Opportunities Act in the U.S. – legally prohibit the use of variables that can lead to an unfair discrimination for specific borrower groups. Prohibited variables usually include variables such as race, color, gender, national origin, and religion. Lenders using digital footprints are likely to face scrutiny whether the digital footprint proxies for such information and therefore violate fair lending acts (see Fuster et al. (2017) on this issue). It is also conceivable that incumbent financial institutions, threatened by competitors using digital footprints, use their well-established access to politicians and regulators to lobby for a restriction of the use of digital footprints on these grounds.

5. Conclusion

In this paper, we have analyzed the information content of the digital footprint – a trail of information that people leave online simply by accessing or registering on a website – for predicting consumer default. Using more than 250,000 observations, we show that even simple, easily accessible variables from the digital footprint match or exceed the information content of credit bureau (FICO) scores.

The correlation between the score based on the digital footprint variables and the FICO score is approximately 10%. As a consequence, the discriminatory power of a model using both the FICO score and the digital footprint variables significantly exceeds models that only use the FICO score or only use the digital footprint variables. This suggests that the digital footprint complements rather than substitutes for credit bureau information and a lender that uses information from both sources (FICO + digital footprint)

can make superior lending decisions compared to lenders that only access one of the two sources of information.

We also show that the discriminatory power for unscorable customers matches the discriminatory power for scorable customers. These results suggest that digital footprints can indeed help to overcome information asymmetries between lenders and borrowers when standard credit bureau information is not available. Digital footprints thus have the potential to boost access to credit to parts of the currently two billion working-age adults worldwide that lack access to services in the formal financial sector, thereby fostering financial inclusion and lowering inequality.

Finally, while consumers might plausibly change their online behavior if digital footprints are widely used for lending decisions, we show that some of the digital footprint variables are clearly costly to manipulate, but, more importantly, such a change in behavior can lead to a situation where the use of digital footprints has a considerable impact on everyday life, with consumers constantly considering their digital footprints which are so far usually left without any further thought. Firms and regulators are equally likely to react, with firms managing the digital footprint of their products and regulators scrutinizing compliance with fair lending acts worldwide.

Overall, our results have potentially wide implications for financial intermediaries' business models going forward, for access to credit for the unbanked, and for the behavior of consumers, firms, and regulators in the digital sphere.

Literature

Beck, Demirguc-Kunt, and Honohan (2009): Access to Financial Services: Measurement, Impact, and Policies, *The World Bank Research Observer* 24(1), 119-145.

Belenzon, S., A. K. Chatterji, and B. Daley (2017): Eponymous Entrepreneurs, *American Economic Review* 107(6), 1638-1655.

Berg, T., M. Puri, and J. Rocholl (2017): Loan Officer Incentives, Internal Rating Models and Default rates, Working Paper.

Berger, A., N. Miller, M. Petersen, R. Rajan, and J. Stein (2005): Does Function Follow Organizational Form? Evidence from the Lending Practices of Large and Small Banks, *Journal of Financial Economics* 76(2), 237-269.

Bertrand, M. and E. Kamenica (2017): Coming apart? Lives of the rich and poor over time in the United States, Working Paper.

Boot, A.W. (1999): Relationship Banking: What Do We Know?, *Journal of Financial Intermediation* 9, 7-25.

Boot, A.W. and A.V. Thakor (2000): Can Relationship Banking Survive Competition?, *Journal of Finance* 55(2), 679-713.

Brown, M., T. Jappelli, and M. Pagano (2009): Information Sharing and Credit: Firm-level Evidence from Transition Countries, *Journal of Financial Intermediation* 18(2), 151-172.

DeLong, E., D. DeLong, and L. Clarke-Pearson (1988): Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, *Biometrics* 44(3), 837-845.

Diamond, D.W. (1984): Financial Intermediation and Delegated Monitoring, *The Review of Economic Studies* 51 (3), 393–414.

Dietterich, T.G. (1998) : Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, *Neural Computation 10(7), 1895-1923.*

Djankov, S., C. McLiesh, and A. Shleifer (2007): Private credit in 129 countries, *Journal of Financial Economics* 84(2), 299-329.

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

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2017): Predictably Unequal? The Effects of Machine Learning on Credit Markets, Working Paper.

Fuster, A., M. Plosser, P. Schnabl, and J. Vickery (2018): The Role of Technology in Mortgage Lending, Working Paper.

Gao, Q., M. Lin, and R. Sias (2017): Word Matters: The Role of Texts in Online Credit Markets, Working Paper.

Guzman, J., and S. Stern (2016): The State of American Entrepreneurship: New Estimates of the Quantity and Quality of Entrepreneurship for 15 US States, 1988-2014. Working Paper.

Hanley, J. and B. McNeil (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* 143(1), 29-36.

Hertzberg, A., A. Liberman, and D. Paravisini (2016): Adverse Selection on Maturity: Evidence from On-Line Consumer Credit, Working Paper.

Hildebrandt, T., M. Puri, and J. Rocholl (2017): Adverse Incentives in Crowdfunding. *Management Science* 63(3), 587-608.

Iyer, R., A. Khwaja, E. Luttmer, and K. Shue (2016): Screening Peers Softly: Inferring the Quality of Small Borrowers, *Management Science* 62(6), 1554-1577.

Japelli, T. and M. Pagano (1993): Information Sharing in Credit Markets, *Journal of Finance* 48(5), 1693-1718.

Kawai, K., K. Onishi, and K. Uetake (2016): Signaling in Online Credit Markets, Working Paper, Yale University.

Lin, M., N. Prabhala, and S. Viswanathan (2013): Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending, *Management Science* 59(1), 17-35.

Lucas, R. (1976): Econometric Policy Evaluation: A Critique, Carnegie-Rochester Conference Series on Public Policy 1, 19-46.

Mester, L., L. Nakamura, and M. Renault (2007): Transaction accounts and loan monitoring. *Review of Financial Studies* 20, 529-556.

Norden, L. and M. Weber (2010): Credit Line Usage, Checking Account Activity, and Default Risk of Bank Borrowers. *Review of Financial Studies* 23, 3665-3699

Petersen, M., and R. Rajan (1994): The Benefits of Lending Relationships: Evidence from Small Business Data, *Journal of Finance* 49(1), 3-37.

Petersen, M.A and R. Rajan (2002): Does Distance still Matter? The Information Revolution in Small Business Lending, *Journal of Finance* 57(6), 2533-2570.

Puri, M., J. Rocholl, and S. Steffen (2017): What do a million observations have to say about loan defaults? Opening the black box of relationships, *Journal of Financial Intermediation* 31, 1-15.

Rook, D. (1987): The Buying Impulse. *Journal of Consumer Research* 14(2), 189-199.

Spence, M (1973): Job Market Signaling, *Quarterly Journal of Economics* 87(3), 355-374.

Stein, R. (2007): Benchmarking default prediction models: pitfalls and remedies in model validation, *Journal of Risk Model Validation* 1(1), 77-113.

Turkyilmaz, C., E. Erdem, and A. Uslu (2015): The Effects of Personality Traits and Website Quality on Online Impulse Buying, *Procedia - Social and Behavioral Sciences* 175, 98-105.

Vallee, B. and Yao Zeng (2018): Marketplace Lending: A New Banking Paradigm? Working Paper.

Wells, J., V. Parboteeah, and J. Valacich, (2011) : Online Impulse Buying: Understanding The Interplay Between Consumer Impulsiveness and Website Quality. *Journal of The Association for Information Systems*, 12(1), 32-56.

Figure 1a: Number of observations per month

This figure shows the monthly number of observations for scorable customers, for unscorable customers, as well as for the total sample. The sample period is from October 19, 2015 to December 2016. The number of observations for October 2015 is scaled up by a factor of 31/13 to make it comparable to a monthly figure. For variable definitions see Table 1.

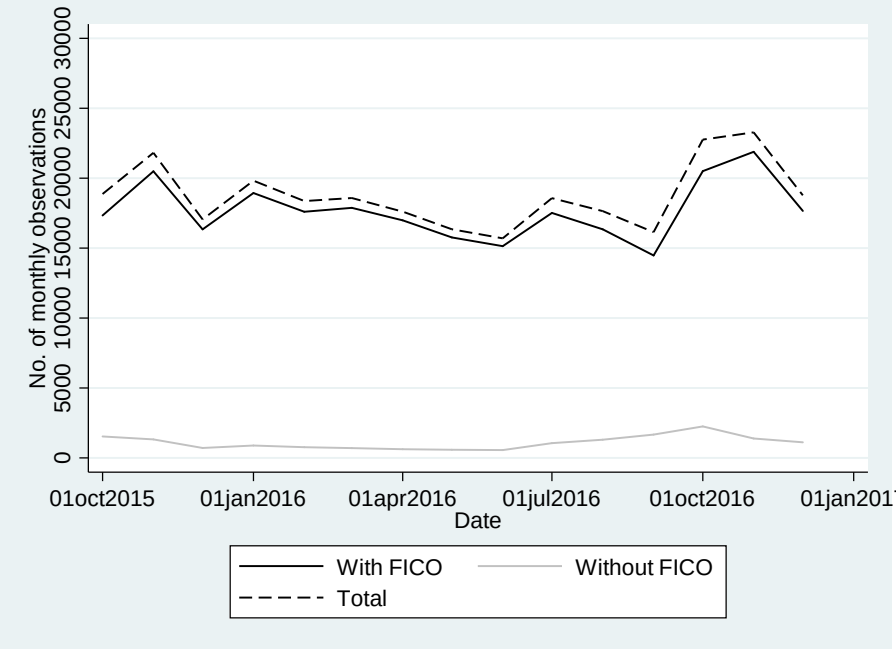

Figure 1b: Geographic distribution of customers in our sample compared to the German population

This figure illustrates the share of customers by state in our sample compared to the German population by state.

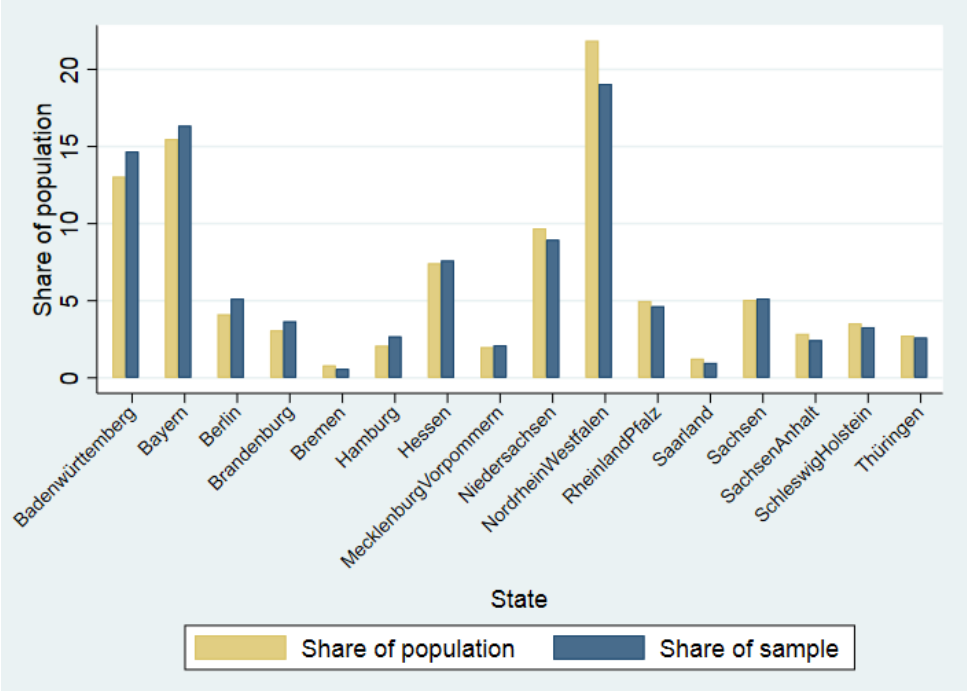

Figure 2: FICO score distribution and default rates

This figure shows the distribution of the FICO score and the raw and smoothed default rates as a function of the FICO score. (*Default(0/1)*) is equal to one if the claim has been transferred to a debt collection agency. The smoothed default rates have been determined using a logistic regression and a second-order polynomial of the FICO score. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

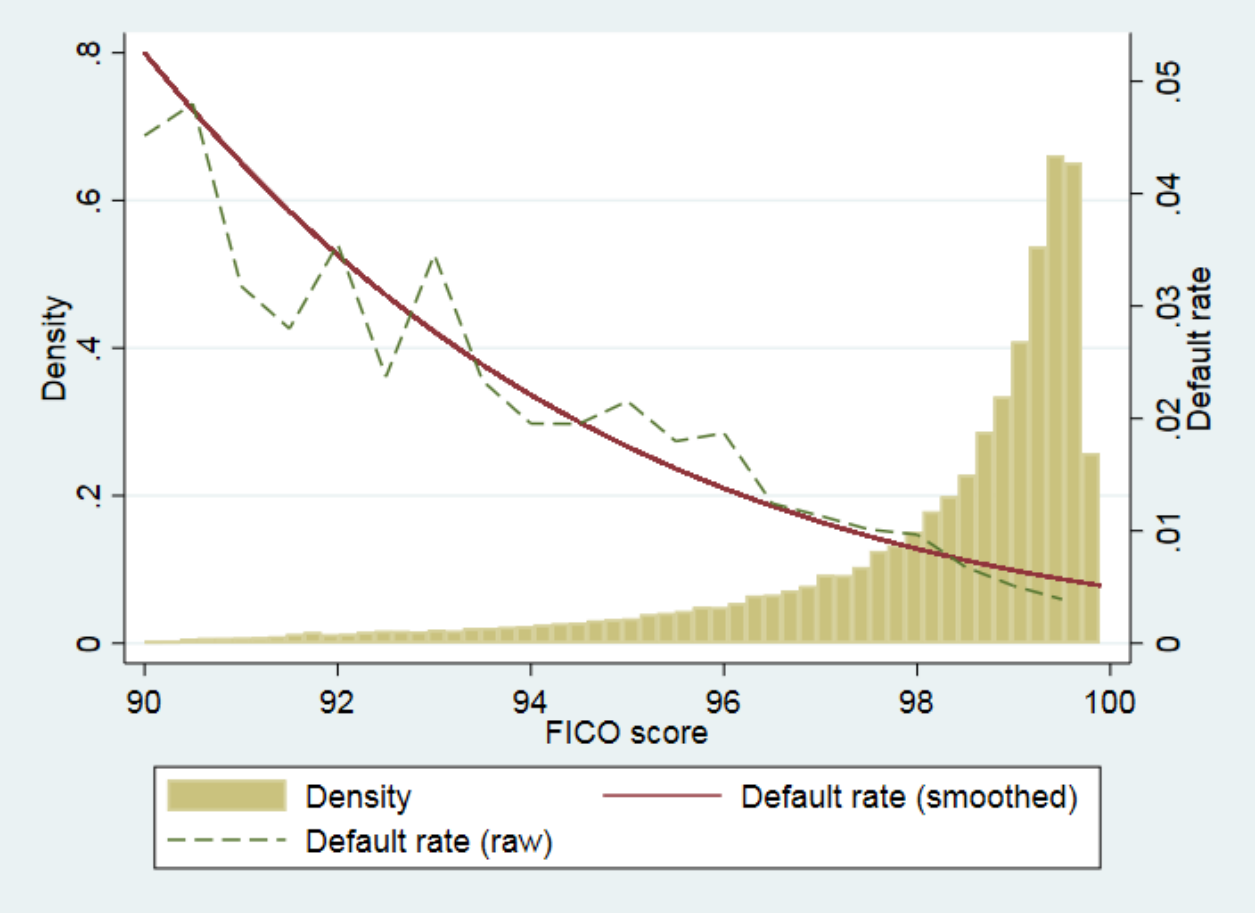

Figure 3: Default rates by combinations of digital footprint variables

This figure shows default rates for combinations of the variables "Operating System" and "Email Host". The x-axis shows default rates, the y-axis illustrates whether the respective dot comes from a single digital footprint variable (for example, "Android users") or whether it comes from a combination of digital footprint variables (for example, "Android + Hotmail"). Default rates for FICO score deciles are provided as reference points in the row at the very bottom. The sample only includes customers with FICO scores. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

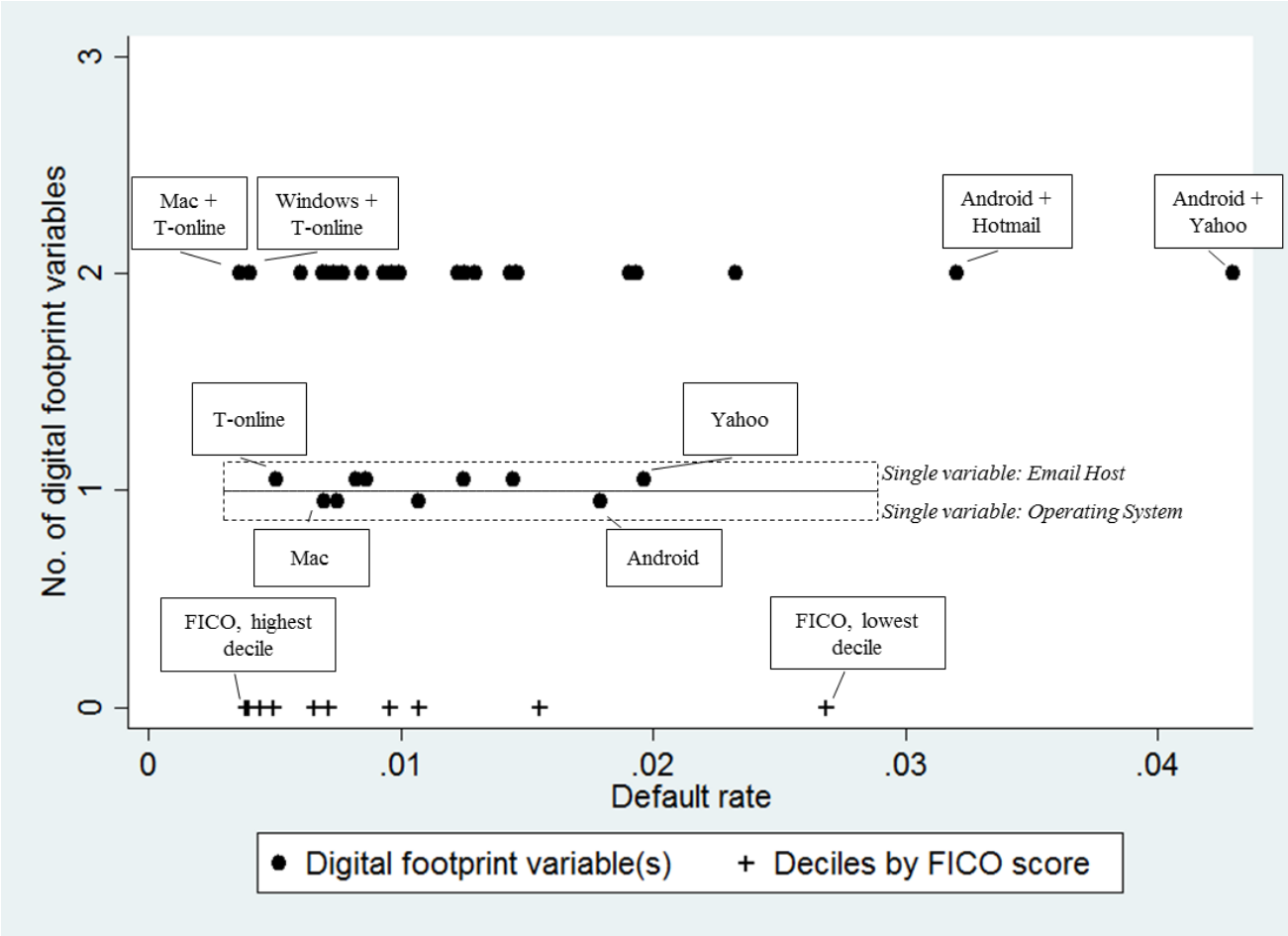

Figure 4: AUC (Area Under Curve) for scorable customers for various model specifications

This figure illustrates the discriminatory power of three different model specifications by providing the receiver operating characteristics curve (ROC-curve) and the area under curve (AUC). The ROC-curves are estimated using a logit regression of the default dummy on the FICO score (light gray), the digital footprint (gray), both FICO and digital footprint (dark gray). The sample only includes customers with FICO scores. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

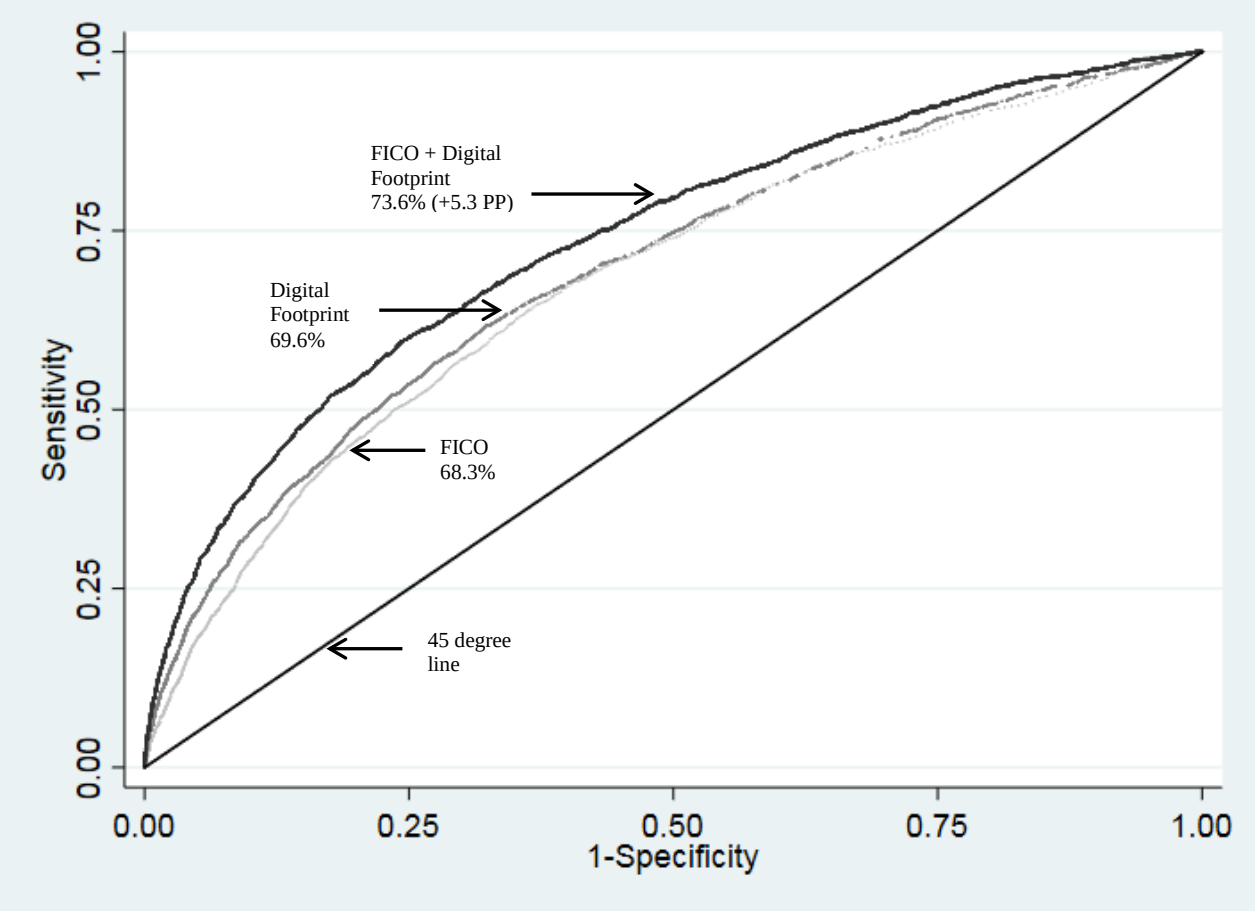

Figure 5: Correlation between Digital Footprint and FICO (scorable customers)

This figure illustrates the correlation between the FICO score and the digital footprint. The x-axis shows percentiles by FICO score. The y-axis shows percentiles by the digital footprint. The digital footprint is estimated using the results from column (2) of Table 5 and multiplied by minus 1 to ensure the same ordering as the FICO (high value = low default probability). The sample only includes customers with FICO scores and is based on a 1% random sample in order to be able to visualize the results. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

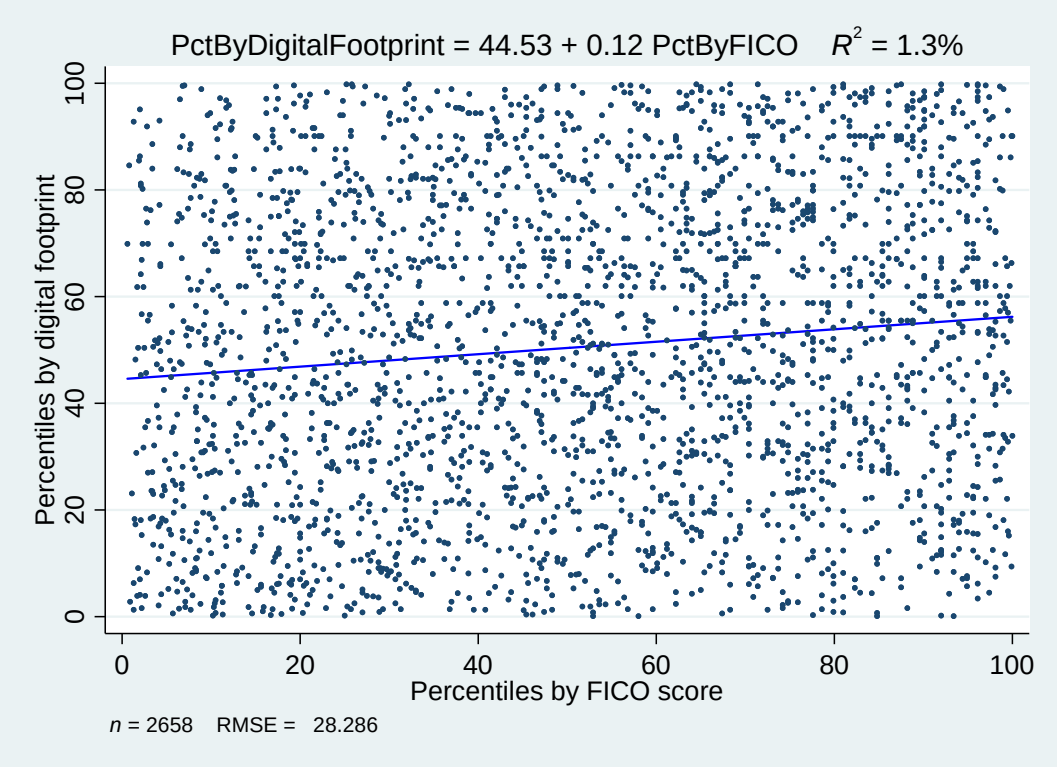

Figure 6: AUC for scorable vs. unscorable customers

This figure illustrates the discriminatory power for different samples by providing the receiver operating characteristics curve (ROC-curve) and the area under curve (AUC) for scorable customers (light gray) and unscorable customers (dark gray). The ROC-curves are estimated using a logistic regression of the default dummy on the digital footprint. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

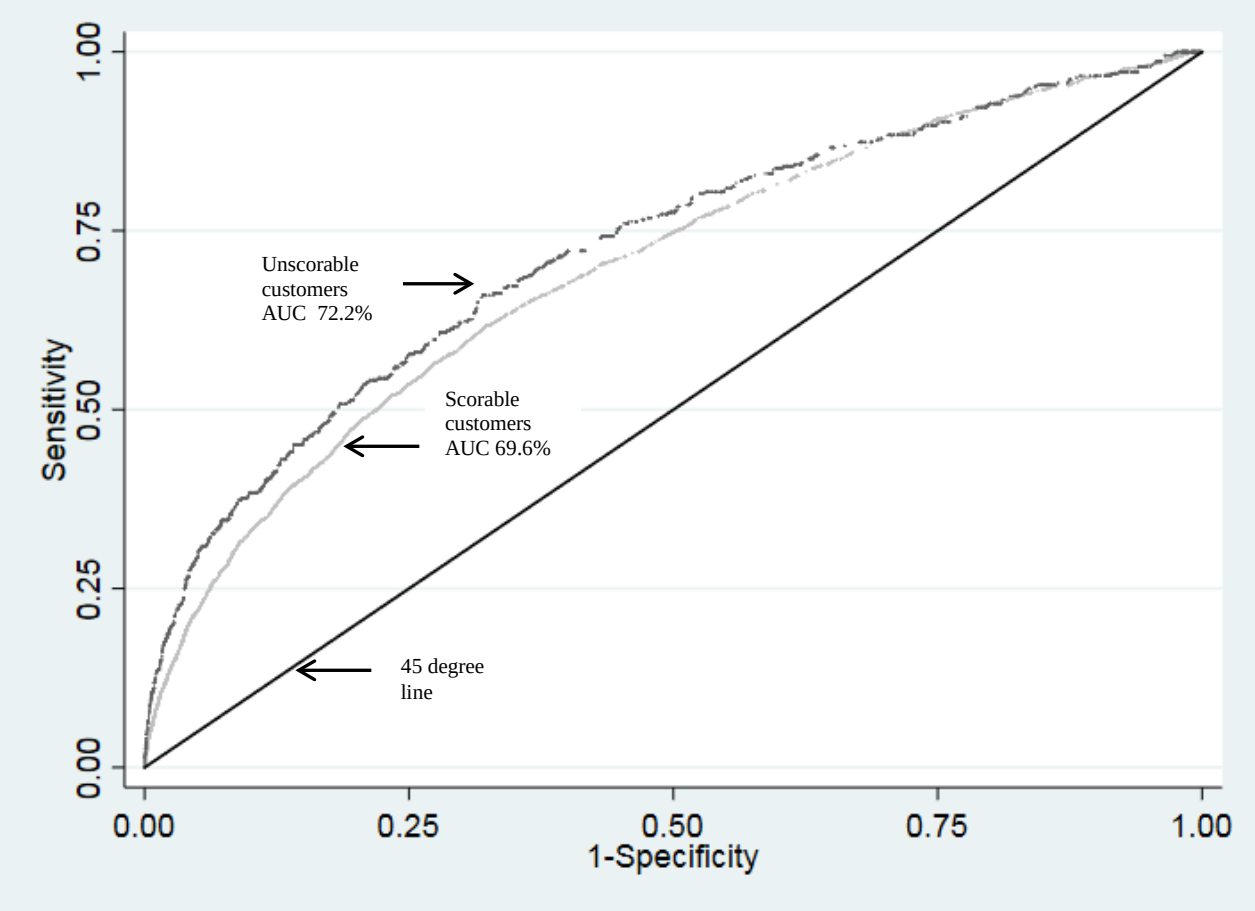

Figure 7: Digital footprint and subsequent changes in the FICO score

This figure illustrates the predictive power of the digital footprint for subsequent changes in the FICO score. The grey line depicts subsequent FICO scores when the creditworthiness using the digital footprint is better than the creditworthiness using the FICO score. The black line depicts subsequent FICO scores when the creditworthiness using the digital footprint is worse than the creditworthiness using FICO. Values are shown by decile of FICO score.

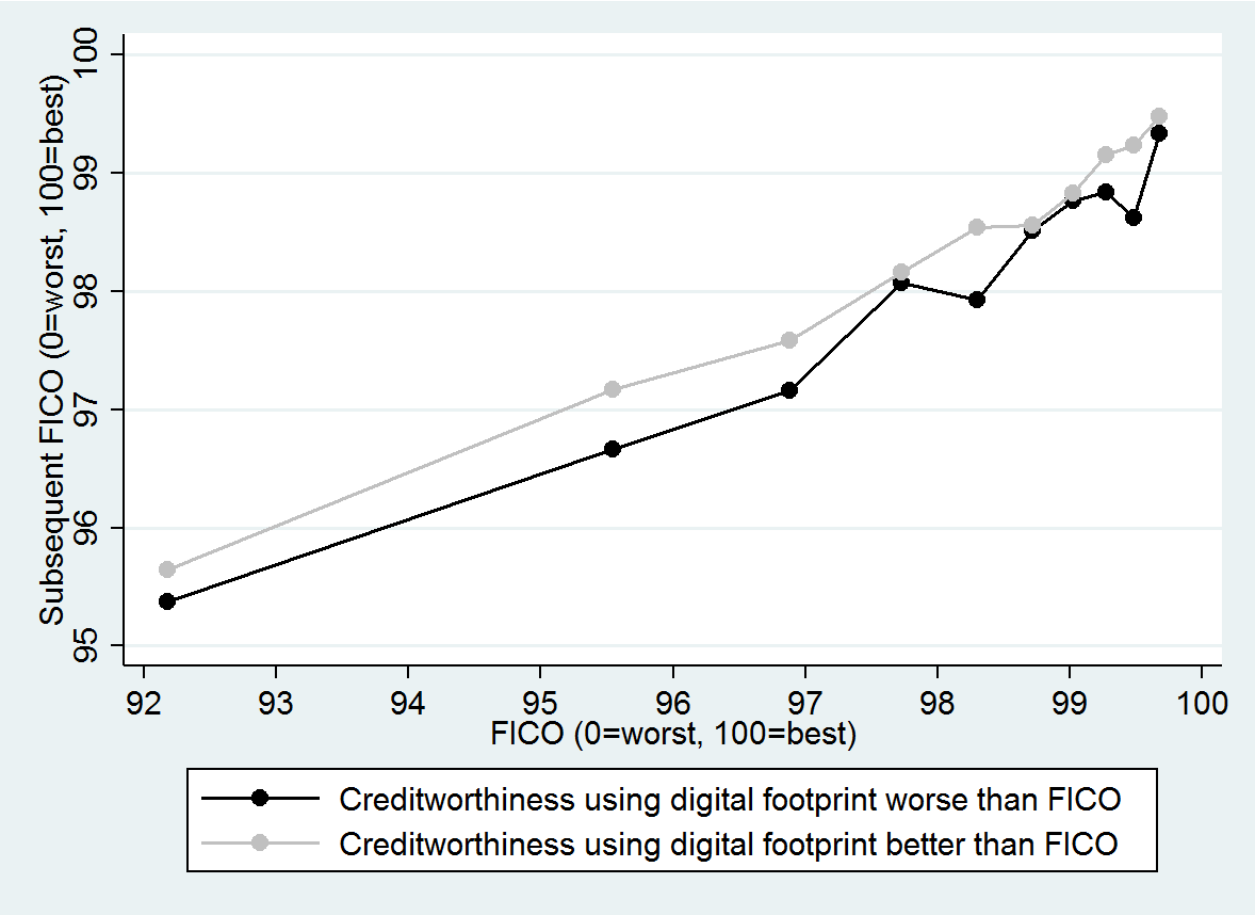

Table 1: Description of variables

Variable	Description	Unit
Order, customer, FICO, and payment behavior		
Order amount	Purchase amount in EUR	Numerical variable
Gender	Gender of customer (female or male)	Dummy variable
Age	Age of customer in years. Information about age is obtained from the credit bureau. Missing information on age indicate that the credit bureau does not have information about a customer's age.	Numerical variable
FICO score	Credit bureau score, similar to the FICO score in the U.S. The score is based on credit history data from various banks, sociodemographic data, as well as payment behavior data sourced from retail sales firms, telecommunication companies, and utilities.	Numerical variable, 0=worst, 100=best
Default	Dummy variable equal to one if the claim is transferred to a debt collection agency (i.e., the customer did not pay the invoice after the third reminder of the firm).	Dummy variable
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
Email Host	Email host. Main examples: Gmx, Web, T-Online, Gmail, Yahoo, Hotmail.	Categorical variable
Channel	Channel through which customer comes to website. Main examples: Paid (including paid and retargeted clicks), Direct, Affiliate, Organic.	Categorical variable
Check-Out Time	Time of day of purchase.	Numerical variable (0-24hrs)
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 Email	Dummy equal to one if first or last name of customer is part of email address.	Dummy variable
Number in Email	Dummy equal to one if a number is part of email address.	Dummy variable
Is Lower Case	Dummy equal to one if first name, last name, street, or city are written in lower case.	Dummy variable
Email Error	Dummy equal to one if email address contains an error in the first trial (Note: Clients can only order if they register with a correct email address).	Dummy variable

Table 2: Descriptive statistics

This table presents summary statistics for the whole sample. The sample period is from October 19, 2015 to December 2016. Panel A provides descriptive statistics for customers with FICO score. Panel B provides descriptive statistics for customers without FICO score. For variable definitions see Table 1.

Panel A: Customers with FICO score							
Variable	**Unit**	**N**	**Mean**	**Std.**	**P25**	**Median**	**P75**
Order and customer							
Order amount	Euro	254,808	317.75	317.10	119.99	218.92	399.98
Gender	Dummy (0=male, 1=female)	254,808	0.66	0.47	0	1	1
Age[#]	Number	254,604	45.06	13.31	34	45	54
FICO score	Number (0=worst, 100=best)	254,808	98.11	2.05	97.58	98.86	99.41
Payment behavior							
Default	Dummy (0/1)	254,808	0.009	0.096	0	0	0
Panel B: Customers without FICO score							
Variable	**Unit**	**N**	**Mean**	**Std.**	**P25**	**Median**	**P75**
Order and customer							
Order amount	Euro	15,591	324.58	319.26	119.99	220.99	399.99
Gender	Dummy (0=male, 1=female)	15,591	0.70	0.46	0	1	1
Age[#]	Number	555	38.20	10.46	30	35	46
FICO score	Number (0=worst, 100=best)	15,591	n.a.	n.a.	n.a.	n.a.	n.a.
Payment behavior							
Default	Dummy (0/1)	15,591	0.025	0.156	0	0	0

[#] Based on information from the credit bureau. Missing information on age indicate that the credit bureau does not have information about a customer's age. Observations with non-missing age in Panel B are cases where the credit bureau has information about the age of the customer, but not enough information to provide a credit score.

Table 3: FICO score, digital footprint variables, and default rates (scorable customers)

This table provides default rates by FICO score quintile as well as default rates by category of each of the digital footprint variables. The sample is based on scorable customers, i.e. the set of customers for which a FICO score is available. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

Variable	Value	Observations	Proportion	Default rate	T-test against Baseline
FICO Score	All	254,808	100%	0.94%	
(by quintile)	Q1 - lowest	50,966	20%	2.12%	Baseline
	Q2	50,964	20%	1.01% ***	(14.19)
	Q3	50,956	20%	0.68% ***	(19.51)
	Q4	50,983	20%	0.47% ***	(23.33)
	Q5 - highest	50,939	20%	0.39% ***	(24.93)
Device	All	254,808	100%	0.94%	
	Desktop	145,871	57%	0.74%	Baseline
	Tablet	45,575	18%	0.91% ***	(-3.62)
	Mobile	26,808	11%	2.14% ***	(-21.84)
	Do not track setting	36,554	14%	0.88% ***	(-2.90)
Operating System	All	254,808	100%	0.94%	
	Windows	124,598	49%	0.74%	Baseline
	iOS	41,478	16%	1.07% ***	(-6.35)
	Android	29,089	11%	1.79% ***	(-16.64)
	Macintosh	21,162	8%	0.69%	(0.79)
	Other	1,927	1%	0.11% *	(1.74)
	Do not track setting	36,554	14%	0.88% ***	(-2.66)
Email Host	All	254,808	100%	0.94%	
	Gmx (partly paid)	58,609	23%	0.82%	Baseline
	Web (partly paid)	54,864	22%	0.86%	(-0.70)
	T-Online (affluent customers)	30,277	12%	0.51% ***	(5.32)
	Gmail (free)	27,842	11%	1.25% ***	(-6.02)
	Yahoo (free, older service)	11,922	5%	1.96% ***	(-11.33)
	Hotmail (free, older service)	10,241	4%	1.45% ***	(-6.11)
	Other	61,053	24%	0.90%	(-1.38)
Channel	All	254,808	100%	0.94%	
	Paid	111,398	44%	1.11%	Baseline
	Direct	45,178	18%	0.84% ***	(4.77)
	Affiliate	24,769	10%	0.64% ***	(6.68)
	Organic	18,295	7%	0.86% ***	(3.00)
	Other	18,614	7%	0.69% ***	(5.24)
	Do not track setting	36,554	14%	0.88% ***	(3.68)
Check-Out Time	All	254,808	100%	0.94%	
	Evening (6pm-midnight)	108,543	43%	0.85%	Baseline
	Night (midnight-6am)	6,913	3%	1.97% ***	(-9.49)
	Morning (6am-noon)	46,600	18%	1.09% ***	(-4.55)
	Afternoon (noon-6pm)	92,752	36%	0.89%	(-0.91)
Do not track setting	All	254,808	100%	0.94%	
	No	218,254	86%	0.94%	Baseline
	Yes	36,554	14%	0.88%	(1.12)
Name in Email	All	254,808	100%	0.94%	
	No	71,016	28%	1.24%	Baseline
	Yes	183,792	72%	0.82% ***	(9.99)
Number in Email	All	254,808	100%	0.94%	
	No	213,640	84%	0.84%	Baseline
	Yes	41,168	16%	1.41% ***	(-10.95)
Is Lower Case	All	254,808	100%	0.94%	
	No	235,560	92%	0.84%	Baseline
	Yes	19,248	8%	2.14% ***	(-18.07)
Email Error	All	254,808	100%	0.94%	
	No	251,308	99%	0.88%	Baseline
	Yes	3,500	1%	5.09% ***	(-25.71)

Table 4: Correlation/Association between FICO score and digital footprint variables (customers with FICO score)

This table provides a measure of association, Cramér's V, between FICO score quintiles and the digital footprint variables. Cramér's V measures the association between two categorical variables and is bounded between [0,1], with 0 denoting no association and 1 denoting perfect association. To allow calculation of Cramér's V, the FICO score has been transformed into categories by forming quintiles of the FICO score. Please note that most digital footprint variables are nominal categorical variables so that Pearson's correlation coefficient or Spearman's rank correlation cannot be determined. The sample is based on scorable customers, i.e. the set of customers for which a FICO score is available. The sample period is from October19, 2015 to December 2016. For variable definitions see Table 1.

	FICO score quintile[1]	Device Type	Operating System	Email Host	Channel	Check-Out Time[1]	Name in Email	Number in Email	Is Lower Case	Email Error
FICO score quintile[1]	1.00***	0.07***	0.05***	0.07***	0.03***	0.03***	0.01***	0.07***	0.02***	0.01
Device Type		1.00***	0.71***[2]	0.07***	0.06***[2]	0.04***	0.05***	0.06***	0.07***	0.01***
Operating System			1.00***	0.08***	0.06***[2]	0.04***	0.06***	0.08***	0.06***	0.01***
Email 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***
Check-Out Time[1]						1.00***	0.01***	0.01***	0.01***	0.01*
Name in Email							1.00***	0.22***	0.01***	0.02***
Number in Email								1.00***	0.02***	0.00**
Is Lower Case									1.00***	0.03***
Email Error										1.00***

1. Transformed into quintiles.
2. We exclude customers with a do not track setting, as the setting simultaneously applies to device, operating system, and channel information.

Table 5: Default regressions (scorable customers)

We estimate default rate regressions where the dependent variable (*Default(0/1)*) is equal to one if the claim has been transferred to a debt collection agency. Column (1) provides results using the FICO score as independent variable, column (2) provides results using the digital footprint variables as independent variables, column (3) uses both the FICO score and the digital footprint variables as independent variables, and column (4) adds age and month and region fixed effects. All models are estimated using a logistic regression model. The sample is based on scorable customers, i.e. the set of customers for which a FICO score is available. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

	(1)		(2)		(3)		(4)	
	FICO		Digital footprint		FICO & digital footprint		FICO & Digital footprint, fixed effects	
VARIABLES	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat
FICO score	-0.17***	(-7.89)			-0.15***	(-6.67)	-0.14***	(-5.83)
Computer & Operating system								
Desktop/Windows			Baseline		Baseline		Baseline	
Desktop/Macintosh			-0.07	(-0.53)	-0.13	(-1.03)	-0.17	(-1.35)
Tablet/Android			0.29***	(3.19)	0.29***	(3.06)	0.30***	(3.25)
Tablet/iOS			0.08	(1.05)	0.08	(0.97)	0.06	(0.81)
Mobile/Android			1.05***	(17.26)	0.95***	(15.34)	0.97***	(15.53)
Mobile/iOS			0.72***	(9.07)	0.57***	(6.73)	0.56***	(6.65)
Email Host [26]								
Gmx (partly paid)			Baseline		Baseline		Baseline	
Web (partly paid)			-0.00	(-0.01)	-0.02	(-0.22)	-0.01	(-0.07)
T-Online (affluent customers)			-0.40***	(-3.89)	-0.35***	(-3.34)	-0.27**	(-2.53)
Gmail (free)			0.34***	(3.79)	0.28***	(3.08)	0.28***	(2.98)
Yahoo (free, older service)			0.75***	(9.19)	0.72***	(8.98)	0.72***	(8.71)
Hotmail (free, older service)			0.35***	(3.70)	0.28***	(2.73)	0.25**	(2.37)
Channel								
Paid			Baseline		Baseline		Baseline	
Affiliate			-0.49***	(-5.33)	-0.54***	(-5.56)	-0.56***	(-5.67)
Direct			-0.27***	(-4.24)	-0.28***	(-4.43)	-0.19***	(-3.10)
Organic			-0.15*	(-1.79)	-0.15*	(-1.73)	-0.13	(-1.52)
Check-Out Time								
Evening (6pm-midnight)			Baseline		Baseline		Baseline	
Morning (6am-noon)			0.28***	(4.52)	0.28***	(4.62)	0.29***	(4.77)
Afternoon (noon-6pm)			0.08	(1.42)	0.08	(1.47)	0.09*	(1.77)
Night (midnight-6am)			0.80***	(7.74)	0.75***	(7.11)	0.73***	(6.86)
Do not track setting			-0.02	(-0.25)	-0.07	(-0.90)	-0.05	(-0.63)
Name In Email			-0.28***	(-5.67)	-0.29***	(-5.69)	-0.29***	(-5.81)
Number In Email			0.26***	(4.50)	0.23***	(3.92)	0.22***	(3.82)
Is Lower Case			0.76***	(13.04)	0.74***	(13.16)	0.74***	(13.25)
Email Error			1.66***	(20.01)	1.67***	(20.37)	1.70***	(20.56)
Constant	12.43***	(5.77)	-4.92***	(-62.84)	9.97***	(4.49)	10.46***	(4.76)
Control for age	No		No		No		Yes	
Month & region fixed effects	No		No		No		Yes	
Observations	254,808		254,808		254,808		254,604	
Pseudo R^2	**0.0244**		**0.0525**		**0.0718**		**0.0808**	
AUC	**0.683**		**0.696**		**0.736**		**0.749**	
(SE)	**(0.006)**		**(0.006)**		**(0.005)**		**(0.005)**	
Difference to AUC=50%	**0.183*** **		**0.196*** **		**0.236*** **		**0.249*** **	
Difference AUC to (1)			**0.013***		**0.053*** **		**0.067*** **	

[26] We only report coefficients for the 6 largest email providers even though we use the largest 18 categories in the regression (all email providers with at least 1000 observations). Using only the 6 reported email hosts does not affect our results.

Table 6: FICO score, digital footprint variables, and default rates (unscorable customers)

This table provides default rates by FICO score quintile as well as default rates by category of each of the digital footprint variables. The sample is based on *unscorable* customers, i.e. the set of customers for which a FICO score is *not* available. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

Variable	Value	Observations	Proportion	Default rate	T-test against baseline
Device	All	15,591	100%	2.49%	
	Desktop	9,191	59%	2.15%	Baseline
	Tablet	2,618	17%	1.64%	(1.63)
	Mobile	1,546	10%	6.21% ***	(-9.07)
	Do not track setting	2,236	14%	2.28%	(-0.37)
Operating System	All	15,591	100%	2.49%	
	Windows	7,770	50%	2.19%	Baseline
	iOS	2,424	16%	2.35%	(-0.48)
	Android	1,646	11%	4.80% ***	(-6.01)
	Macintosh	1,421	9%	1.69%	(1.20)
	Other	94	1%	7.44% ***	(-3.42)
	Do not track setting	2,236	14%	2.28%	(-0.26)
Email Host	All	15,591	100%	2.49%	
	Gmx (partly paid)	3,681	24%	2.42%	Baseline
	Web (partly paid)	3,352	21%	2.63%	(-0.55)
	T-Online (affluent customers)	1,711	11%	1.52% **	(2.13)
	Gmail (free)	1,694	11%	3.60% **	(-2.44)
	Yahoo (free, older service)	732	5%	3.14%	(-1.14)
	Hotmail (free, older service)	546	4%	2.75%	(-0.46)
	Other	3,875	25%	2.22%	(0.57)
Channel	All	15,591	100%	2.49%	
	Paid	6,447	41%	2.89%	Baseline
	Direct	3,262	21%	1.87% ***	(3.00)
	Affiliate	1,395	9%	2.65%	(0.47)
	Organic	1,178	8%	2.55%	(0.64)
	Other	1,073	7%	2.14%	(1.36)
	Do not track setting	2,236	14%	2.28% ***	(1.51)
Check-Out Time	All	15,591	100%	2.49%	
	Evening (6pm-midnight)	6,349	41%	2.05%	Baseline
	Night (midnight-6am)	369	2%	3.52% **	(-1.91)
	Morning (6am-noon)	2,960	19%	2.74% **	(-2.08)
	Afternoon (noon-6pm)	5,913	38%	2.77% ***	(-2.63)
Do not track setting	All	15,591	100%	2.49%	
	No	13,355	86%	2.52%	Baseline
	Yes	2,236	14%	2.28%	(0.68)
Name in Email	All	15,591	100%	2.49%	
	No	4,433	28%	3.93%	Baseline
	Yes	11,158	72%	1.92% ***	(7.27)
Number in Email	All	15,591	100%	2.49%	
	No	12,967	83%	1.99%	Baseline
	Yes	2,624	17%	4.95% ***	(-8.91)
Is Lower Case	All	15,591	100%	2.49%	
	No	14,566	93%	2.21%	Baseline
	Yes	1,025	7%	6.44% ***	(-8.42)
Email Error	All	15,591	100%	2.49%	
	No	15,305	98%	2.31%	Baseline
	Yes	286	2%	12.24% ***	(-10.72)

Table 7: Default regressions (unscorable customers)

This table provides regression results for the same model specifications as in Table 5 using the sample of *unscorable* customers. Since this sample only includes customers for whom FICO scores are unavailable, the first column from the initial specification is omitted. We estimate default rate regressions where the dependent variable (*Default(0/1)*) is equal to one if the claim has been transferred to a debt collection agency. Column (1) provides results using the digital footprint variables as independent variables. Column (2) provides a comparison to the results for the sample of scorable customers. Column (3) adds month and region fixed effects. All models are estimated using a logistic regression model. The sample is based on *unscorable* customers, i.e. the set of customers for which a FICO score is *not* available. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

	(1) Digital footprint for unscorable customers		(2) For comparison: Digital footprint for scorable customers (column (2) of Table 5)		(3) Digital footprint for unscorable customers, fixed effects	
VARIABLES	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat
Computer & Operating system						
Desktop/Windows	Baseline		Baseline		Baseline	
Desktop/Macintosh	-0.26	(-1.10)	-0.07	(-0.53)	-0.22	(-0.93)
Tablet/Android	-0.22	(-0.85)	0.29***	(3.19)	-0.18	(-0.72)
Tablet/iOS	-0.45*	(-1.71)	0.08	(1.05)	-0.47*	(-1.76)
Mobile/Android	1.07***	(5.98)	1.05***	(17.26)	1.05***	(5.56)
Mobile/iOS	0.63***	(2.69)	0.72***	(9.07)	0.60**	(2.42)
Email Host [27]						
Gmx	Baseline		Baseline		Baseline	
Web	0.02	(0.11)	-0.00	(-0.01)	0.02	(0.13)
T-Online	-0.39	(-1.14)	-0.40***	(-3.89)	-0.44	(-1.29)
Gmail	0.32	(1.36)	0.34***	(3.79)	0.32	(1.35)
Yahoo	0.17	(0.61)	0.75***	(9.19)	0.14	(0.45)
Hotmail	-0.02	(-0.06)	0.35***	(3.70)	-0.09	(-0.31)
Channel						
Paid	Baseline		Baseline		Baseline	
Affiliate	-0.08	(-0.39)	-0.49***	(-5.33)	-0.02	(-0.11)
Direct	-0.42**	(-2.35)	-0.27***	(-4.24)	-0.39**	(-2.08)
Organic	-0.05	(-0.24)	-0.15*	(-1.79)	0.03	(0.12)
Check-Out Time						
Evening (6pm-midnight)	Baseline		Baseline		Baseline	
Morning (6am-noon)	0.30*	(1.81)	0.28***	(4.52)	0.29*	(1.74)
Afternoon (noon-6pm)	0.39***	(2.71)	0.08	(1.42)	0.38***	(2.63)
Night (midnight-6am)	0.45	(1.38)	0.80***	(7.74)	0.43	(1.36)
Do not track setting	-0.16	(-0.83)	-0.02	(-0.25)	-0.12	(-0.59)
Name In Email	-0.59***	(-4.68)	-0.28***	(-5.67)	-0.55***	(-4.33)
Number In Email	0.63***	(4.32)	0.26***	(4.50)	0.64***	(4.35)
Is Lower Case	0.95***	(5.44)	0.76***	(13.04)	0.90***	(4.78)
Email Error	1.66***	(7.82)	1.66***	(20.01)	1.67***	(7.33)
Constant	-3.80***	(-19.20)	-4.92***	(-62.84)	-4.11***	(-14.33)
Month & region fixed effects	No		No		Yes	
Observations	15,591		254,808		15,591	
Pseudo R^2	**0.0907**		**0.0525**		**0.139**	
AUC	**0.722**		**0.696**		**0.777**	
(SE)	**(0.014)**		**(0.006)**		**(0.012)**	
Difference to AUC=50%	**0.222*** **		**0.196*** **		**0.277*** **	

[27] We only report coefficients for the 6 largest email providers even though we use the largest 18 categories in the regression (all email providers with at least 1000 observations). Using only the 6 reported email hosts does not significantly affect the results.

Table 8: Out-of-sample estimates

This table provides robustness tests out-of-sample for all main regression specifications. Panel 1 reports AUCs for scorable customers for the model specifications from Table 5, and Panel 2 reports AUCs for unscorable customers for the model specifications from Table 7. Column (1) reports the baseline results for scorable and unscorable customers. Column (2) reports out-of-sample estimates of the AUC using Nx2-fold cross validation. We thereby randomly divide the full sample into half samples A and B. We then estimate a predictive logistic regression using sample A and use the coefficients to create predicted values for observations in sample B. We also estimate a predictive regression using sample B and use the coefficients to create predicted values for observations in sample A. We then determine the AUC for the full sample of observations, using all predicted values estimated out-of-sample. The AUCs reported in Column (2) are the mean AUCs from 100 iterations. The sample period is from October 19, 2015 to December 2016.

	(1) Baseline (In-sample)	(2) Out-of-sample
Panel 1: Scorable customers		
AUC FICO Score	0.683	0.680
N	254,808	254,808
AUC Digital Footprint	0.696	0.688
N	254,808	254,808
AUC FICO + Digital Footprint	0.736	0.728
N	254,808	254,808
AUC FICO + Digital Footprint, fixed effects	0.749	0.721
N	254,604	254,604
Panel 2: Unscorable customers		
AUC Digital Footprint	0.722	0.683
N	15,591	15,591
AUC Digital Footprint, fixed effects	0.777	0.647
N	15,591	15,591

Table 9: Robustness tests (scorable customers)

This table provides robustness tests using alternative default definitions as well as various sample splits. Panel A provides results using alternative default definitions. Column (1) reports results using the standard default definition (default = transfer to debt collection agency), column (2) provides a stricter default definition (default = no full repayment after attempts of debt collection agency). Column (3) uses only the sample of defaulted loans and uses the loss given default as the dependent variable. In column (3) we report the R-squared instead of the AUC, as column (3) is estimated using a linear regression model while all other models are estimated using a logistic regression. Panel B provides results for various sample splits. All models are estimated using a logistic regression model; apart from column (3) in Panel A, which is estimated using a linear regression model. The sample is based on scorable customers, i.e. the set of customers for which a FICO score is available. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

Panel A: Default definition	(1) Baseline (Default = Transfer to collection agency)	(2) Default = Writedown	(3) Loss given default (R^2 reported)
AUC FICO Score	0.6826	0.6918	0.0126
AUC Digital footprint	0.6960	0.7232	0.0650
AUC FICO + Digital Footprint	0.7360	0.7564	0.0715
N	254,808	254,808	2,384

Panel B: Sample splits	(1) Small orders < EUR 218.92	(2) Large orders ≥ EUR 218.92	(3) Female	(4) Male
AUC FICO Score	0.6878	0.6784	0.6893	0.6696
AUC Digital footprint	0.7126	0.6910	0.6997	0.6999
AUC FICO + Digital Footprint	0.7497	0.7306	0.7448	0.7245
N	127,404	127,404	168,366	86,442

Table 10: Predicting changes in FICO scores with the digital footprint

This table provides a linear regression of changes in FICO scores on the difference between the default probability using the digital footprint and the default probability using the FICO score. The independent variable $\Delta(DigitalFootprint_t, FICO_t)$ measures the difference in predicted values of column (2) of Table 5 and the predicted values of column (1) of Table 5. The dependent variable (Δ $(FICO_{t+1}, FICO_t)$) measures the change in FICO score between i) the FICO score as of the date of purchase from Table 5 and ii) the latest available FICO score up to March 2018. Two FICO scores at two different dates are only available when customers return to the E-Commerce company at least once between October 2015 and March 2018. Column (2) adds the initial FICO score as a control variable, column (3) adds month and region fixed effects, column (4) displays results by quintile of $\Delta(DigitalFootprint_t, FICO_t)$, and columns (5) and (6) split the sample at the median FICO score. For variable definitions see Table 1.

Sample	(1) All	(2) All	(3) All	(4) All	(5) $FICO_t$ < Median	(6) $FICO_t \geq$ Median
Dependent variable	Δ ($FICO_{t+1}$, $FICO_t$)	Δ ($FICO_{t+1}$, $FICO_t$)	Δ ($FICO_{t+1}$, $FICO_t$)	Δ ($FICO_{t+1}$, $FICO_t$)	Δ ($FICO_{t+1}$, $FICO_t$)	Δ ($FICO_{t+1}$, $FICO_t$)
Δ ($DigitalFootprint_t$, $FICO_t$)	-74.56***	-28.14***	-29.74***		-34.24***	-20.75**
	(-11.71)	(-4.56)	(-4.95)		(-4.23)	(-2.48)
Q1 (-100% to -0.49%)				0.39**		
				(2.38)		
Q2 (-0.49% to -0.25%)				0.15*		
				(1.74)		
Q3 (-0.25% to -0.05%)				baseline		
Q4 (-0.05% to +0.35%)				0.08		
				(0.92)		
Q5 (+0.35% to +100%)				-0.39***		
				(-3.05)		
$FICO_t$		-0.43***	-0.42***	-0.42***	-0.42***	-0.19
		(-13.81)	(-13.66)	(-10.27)	(-8.38)	(-1.37)
Constant	0.37***	42.31***	absorbed	absorbed	absorbed	absorbed
	(8.64)	(13.84)				
Month & region fixed effects	No	No	Yes	Yes	Yes	Yes
Observations	17,645	17,645	17,645	17,645	8,853	8,792
Adj. R^2	2.74%	6.95%	7.95%	7.92%	7.13%	1.54%

Appendix

Table A.1: Comparability of default rates to other retail data sets

Study	Sample	Default rate	Time horizon	Default rate (annualized)
This study				
This study	270,399 purchases at a German E-Commerce company between October 2015 and December 2016	1.0%	~4 months	3.0%
Germany				
Berg, Puri, and Rocholl (2017)	100,000 consumer loans at a large German private bank, 2008-2010	2.5%	12 months	2.5%
Puri, Rocholl, and Steffen (2017)	1 million consumer loans at 296 German savings banks, 2004-2008	1.1%	12 months	1.1%
Schufa (2017)[1] – study by the major credit bureau in Germany	17.4 million consumer loans covered by the main credit bureau in Germany in 2016	2.2%	12 months	2.2%
Schufa (2016)[1]– study by the major credit bureau in Germany	17.3 million consumer loans covered by the main credit bureau in Germany in 2015	2.4%	12 months	2.4%
Deutsche Bank (2016)[2]	All retail loans of Deutsche Bank (i.e, the largest German bank)	1.5% (Basel II PD estimate)	12 months	1.5%
Commerzbank (2016)[3]	All retail loans of Commerzbank (i.e., the second largest German bank)	2.0% (Basel II PD estimate)	12 months	2.0%
United States				
Federal reserve[4]	Charge-off rate on consumer loans, Q4/2016	2.09%	12 months (annualized quarterly data)	2.09%
Federal reserve[4]	Charge-off rate on consumer loans, Q4/2015	1.76%	12 months (annualized quarterly data)	1.76%
Hertzberg, Liberman, and Paravisini (2016)	12,091 36-months loans from Lending Club issued between December 2012 and February 2013	9.2%	~26 months	4.2%
Lending Club (own analysis)	375,803 36-month loans from Lending Club issued between October 2015 and December 2016	5.11%	12 months	5.11%
Iyer, Khwaja, Luttmer, and Shue (2016)	17,212 36-months loans from Prosper.com issued between February 2007 and October 2008	30.6%	36 months	10.2%
Puri, Hildebrandt, and Rocholl (2017)	12,183 loans from Prosper.com between February 2007- April 2008	10.8%-18.6%	per 1,000 days	3.9%-6.8%

1. Schufa is the main credit bureau in Germany, comparable to Fair Isaac Newton in the U.S., for example, For data on 2016 default rates see Figure 2.11 on page 17 in https://www.schufa.de/media/editorial/themenportal/kredit_kompass_2017/SCHUFA_Kredit-Kompass_2017_neu.pdf . For the data on 2016 default rates see Figure 2.11 on page 18 in https://www.schufa.de/media/editorial/ueber_uns/bilder/studien_und_publikationen/kredit_kompass/SCHUFA_Kredit-Kompass-2016.pdf (available in German only).
2. See Table on page 90 in Deutsche Bank's Pillar 3 Report 2016, available via https://www.db.com/ir/en/download/Deutsche_Bank_Pillar_3_Report_2016.pdf
3. See Table 12 on page 34 in Commerzbank's Disclosure Report 2016, available via https://www.commerzbank.de/media/aktionaere/service/archive/konzern/2017/Disclosure_Report_2016.pdf
4. Series "CORCABAS" in FRED, see https://fred.stlouisfed.org/series/CORCACBS

Table A.2: Comparability of Area-Under-Curve to other retail data sets

Study	Sample	AUC using FICO
Area Under the Curve (AUC) using the FICO score (or comparable credit score) only		
This study	270,399 purchases at a German E-Commerce company in 2015/2016	68.3%
Berg, Puri, and Rocholl (2017)#	100,000 consumer loans at a large German private bank, 2008-2010	66.6%
Puri, Rocholl, and Steffen (2017)#	1 million consumer loans at 296 German savings banks, 2004-2008	66.5%
Iyer, Khwaja, Luttmer, and Shue (2016)	17,212 36-months loans from Prosper.com issued between February 2007 and October 2008	62.5%
Lending Club (own analysis)	375,803 36-month loans from Lending Club issued between October 2015 and December 2016[1]	59.8%

AUC and changes in the Area Under the Curve using other variables in addition to the FICO score			
		AUC Change	Combined AUC
This study	Digital footprint versus FICO score only	+ 5.3PP	73.6%
Berg, Puri, and Rocholl (2017)#	Bank internal rating (which includes FICO score) versus FICO score only	+8.8PP	75.4%
Puri, Rocholl, and Steffen (2017)#	Bank internal rating (which includes FICO score) versus FICO score only	+11.9PP	78.4%
Iyer, Khwaja, Luttmer, and Shue (2016)	Interest rates versus FICO score only	+5.7PP	68.2%
Iyer, Khwaja, Luttmer, and Shue (2016)	All available financial and coded information (including FICO score) versus FICO score only	+8.9PP	71.4%
Lending Club (own analysis)	Lending Club loan grade (which includes FICO score) versus FICO score only	+11.9PP	71.7%

These results are not in the original papers but were provided to us by the authors using exactly the same data set from the paper.
1. Results are very similar for 60-month loans.

Table A.3: Descriptive statistics for computer and operating system category (scorable customers)

This table provides descriptive statistics for the computer and operating system category. The sample is based on scorable customers, i.e. the set of customers for which a FICO score is available. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

Variable	Value	Observations	Proportion	Default rate	T-test against baseline
Computer and Operating system	All	254,808	100%	0.94%	
	Desktop/Windows	123,085	48%	0.74%	Baseline
	Desktop/Macintosh	21,158	8%	0.69%	(0.70)
	Desktop/other	1,628	1%	0.74%	(0.03)
	Tablet/Android	15,111	6%	1.11% ***	(-4.86)
	Tablet/iOS	29,940	12%	0.79%	(-0.88)
	Tablet/other	524	0%	1.53% **	(-2.08)
	Mobile/Android	13,967	5%	2.53% ***	(-20.92)
	Mobile/iOS	11,531	5%	1.80% ***	(-11.90)
	Mobile/other	1,310	1%	1.15% *	(-1.68)
	Do not track setting	36,554	14%	0.88% ***	(-2.70)

Table A.4: Descriptive statistics for computer and operating system category (unscorable customers)

This table provides descriptive statistics for the computer and operating system category. The sample is based on unscorable customers, i.e. the set of customers for which a FICO score is not available. The sample period is from October 19, 2015 to December 2016. For variable definitions see Table 1.

Variable	Value	Observations	Proportion	Default rate	T-test against baseline
Computer and Operating system	All	15,591	100%	2.49%	
	Desktop/Windows	7,688	49%	2.20%	Baseline
	Desktop/Macintosh	1,421	9%	1.69%	(1.22)
	Desktop/other	82	1%	6.10% **	(-2.37)
	Tablet/Android	857	5%	2.10%	(0.19)
	Tablet/iOS	1,737	11%	1.44% **	(2.01)
	Tablet/other	24	0%	0.00%	(0.73)
	Mobile/Android	789	5%	7.73% ***	(-9.15)
	Mobile/iOS	687	4%	4.66% ***	(-4.04)
	Mobile/other	70	0%	4.43%	(-1.18)
	Do not track setting	2,236	14%	2.28%	(-0.23)