You are how you pay: Understanding and identifying the payment behavior of sociodemographic groups

AUREL RUBEN MÄDER, Swiss National Bank & EPFL, Switzerland
MATTHIAS JÜTTNER, Swiss National Bank, Switzerland
DANIEL GATICA-PEREZ, Idiap Research Institute & EPFL, Switzerland

Understanding the payment behavior of sociodemographic groups is important for public institutions in designing inclusive policies. Thus, public institutions regularly conduct payment surveys to monitor the payment behavior of these groups. However, such surveys are costly, conducted infrequently, and limited in the number of participants. This paper presents a methodology that enables policy-makers to monitor the payment behavior of sociodemographic groups with card data while complying with privacy rights. Specifically, it provides a correlational analysis of payment behavior across sociodemographic groups, demonstrates the potential of payment data to infer sociodemographic information, and proposes a methodology for enriching card data with this information. This paper reveals that sociodemographic groups exhibit different payment behaviors, that groups can be inferred from payment data, and that anonymized card data can be enriched with sociodemographic information. The proposed methodology enables public institutions to complement surveys with timely sociodemographic insights from anonymized card data, reducing costs, easing participant burden, and allowing more frequent updates.

CCS Concepts: • **Social and professional topics → Government technology policy**; • **Applied computing → Economics**; • **Human-centered computing →** *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: Card payments, payment behavior, sociodemographics, government policy, machine learning

ACM Reference Format:
Aurel Ruben Mäder, Matthias Jüttner, and Daniel Gatica-Perez. 2026. You are how you pay: Understanding and identifying the payment behavior of sociodemographic groups. 1, 1 (March 2026), 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Public institutions need to understand the payment behavior of sociodemographic groups to adjust their policies to suit all parts of society. Inclusive policies concerning payment systems are especially important given the recent drastic changes in payment behavior. While consumers still predominantly used cash before the global COVID-19 pandemic, they have used electronic payment instruments since the pandemic [4, 13, 44]. However, electronic payment instruments

We thank Martin Brown, Laura Felber, Christoph Meyer, Thomas Nellen, Michael Zimmert, and members of the banking operations analysis and cash circulation teams at the Swiss National Bank for their valuable feedback and advice. This paper uses data provided by the Swiss National Bank, PostFinance Ltd., and Worldline Switzerland Ltd., and would not have been possible without the support of those institutions.
The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of the Swiss National Bank (SNB). The SNB takes no responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper.
Authors' Contact Information: Aurel Ruben Mäder, Swiss National Bank & EPFL, Switzerland, aurel.maeder@snb.ch; Matthias Jüttner, Swiss National Bank, Switzerland; Daniel Gatica-Perez, gatica@idiap.ch, Idiap Research Institute & EPFL, Switzerland.

Manuscript submitted to ACM

are not equally accessible for all population groups. Older and low-income population groups still heavily rely on cash [27, 44]. Thus, some public institutions have started to promote inclusive electronic payment instruments. The Central Bank of Brazil, for instance, established an electronic payment instrument that is designed to be accessible to all parts of society [9, 20]. However, to design inclusive payment instruments, public institutions must first understand the payment behavior of sociodemographic groups.

In the past, public institutions and researchers have used payment surveys to understand the payment behavior of sociodemographic groups [6, 14, 34]. Such surveys offer rich data about payment behavior, preferences, and the associated sociodemographic characteristics of the survey participants. However, payment surveys entail high costs, are conducted infrequently, and are limited in the number of participants. Thus, public institutions have started to use card data as alternative payment data sources [2, 28, 31]. Card data have the advantages of being comparatively cheap, available almost in real time, and having a high data volume. Card data are anonymized given privacy concerns and, thus, lack sociodemographic information. Thus, they cannot currently be used as a data source to consider the needs of different sociodemographic groups.

Using novel datasets consisting of payment surveys and anonymized card payment data, we propose an approach to enrich card data with sociodemographic information while complying with the privacy rights of cardholders. This paper begins by investigating payment behavior across different sociodemographic groups through correlational analysis using payment survey data. We then show that payment survey data can be used to infer sociodemographic variables, with accuracies varying across variables. Finally, to enrich card data with probabilistic sociodemographic information, this paper proposes a methodology that trains machine learning models on survey data and then employs the models on card data. Validation exercises are performed to substantiate the estimated sociodemographic information.

This paper establishes three main findings. First, different sociodemographic groups exhibit different payment behaviors and rely on different payment instruments. People aged 65 and above, for instance, still rely heavily on cash, whereas the rest of the population tends to use more electronic payment instruments. Second, sociodemographic groups can be identified from their payment behavior, with accuracies ranging between 60% and 83%. More specifically, sociodemographic variables, including sex, marital status, and financial status, are inferred with accuracies ranging between 60% and 70%, whereas inferences of variables such as education and age group have accuracies between 70% and 83%. Third, sociodemographic distributions can be estimated from anonymized card data, which match official demographic census data at the postal code level, with correlations ranging between $-0.18R$ and 0.66. Sociodemographic enriched card data suggest that temporal payment patterns depend on the inferred age groups.

This paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 presents the payment and demographic data, and Section 4 outlines the methodology, including feature extraction, statistical and machine learning methods, and the inference pipeline. Section 5 reports a correlational analysis of payment behavior across sociodemographic groups, examines how sociodemographic traits can be inferred from survey and card payment data, and compares these estimates with regional distributions in Switzerland. Section 6 discusses policy implications, privacy considerations, and limitations, and Section 7 concludes the paper.

2 Related literature

This study draws on literature that explores payment behavior, payment data, and the inference of sociodemographic information.

Payment surveys have already established correlations between sociodemographic variables and payment behavior. Surveys have revealed strong correlations between the ownership of certain payment instruments and age [6, 14, 19].

Younger people (below 40) are, for instance, less likely to own a credit card than older people (above 40). Younger, more urban people are, however, more likely to adopt and use digital payment methods such as the contactless feature of payment cards [11]. The levels of income and education are similarly correlated with cash or card usage. While low-income and high-school-educated people tend to use cash, high-income and university-educated people tend to use credit cards [14]. Recent studies show that electronic payment usage has further increased since the COVID-19 pandemic, but only in countries with strong digital infrastructure and with population groups already familiar with such systems [10, 15]. For instance, Brown et al. [10] finds that raising contactless card payment limits during the pandemic mainly boosted spending among existing, but not first-time, users. This highlights the importance of understanding payment behavior across sociodemographic groups, as it helps policymakers assess how changes in the payment landscape, such as declining ATM density or growing financial innovation, affect different population groups [10, 11, 14].

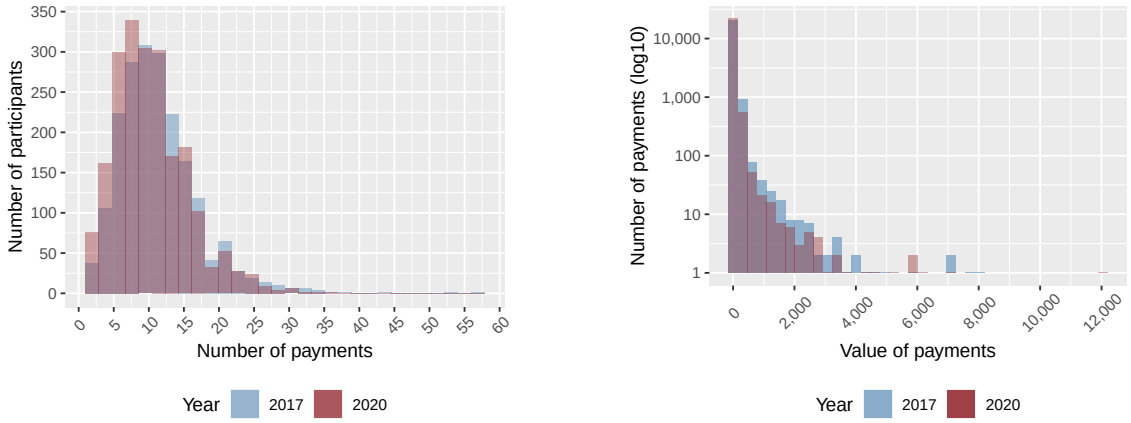
Card data are increasingly used to inform policy decisions. They have been used to measure economic activity and shocks at a fine-grained geographical and sector level in almost real time. Governmental institutions augment their classical methodology of consumer and business surveys by estimating economic metrics with card data. The FED [2], the ECB [28], the SNB [26] and many other organizations [1, 7, 31] use card data to estimate metrics such as GDP, consumption spending and economic shocks. Some studies have explored whether provincial-level card data metrics are correlated with sociodemographic indices. Sobolevsky et al. [40] use aggregated features based on individual credit and debit card transactions to estimate sociodemographic variables at the provincial level. They obtain correlations between 0.3 (crime rate) and 0.64 (education level, life expectancy). Di Clemente et al. [18] show that payment sequences can be used to identify urban groups on the basis of their spending habits. They cluster sequences of payment purposes (e.g., supermarkets, restaurants, and road fees) and find distinct urban groups such as young, high-tech, and commuter groups [18]. The resulting policy-driven field of literature offers economic metrics with the great advantage of higher spatial and temporal resolution while keeping costs lower than those of classical economic surveys.

Studies have used other behavioral data besides card data to infer sociodemographic groups and have discussed the privacy implications of such data. It has been shown that digital traces can be used to identify sociodemographic groups. Kosinski et al. [33] showed that Facebook likes can be used to infer highly personal traits of users. The study collected sociodemographic profiles of 58,000 Facebook users and used the likes of Facebook sites and posts to infer the collected variables. The study yielded accuracies for sociodemographic variables ranging from 65%-70% (relationship status, drug use) to 85%-95% (ethnicity, gender, sexuality). Another study revealed how text search queries can be used to infer the gender, age, religion and political views of internet users [8]. Alizadeh et al. [3] show that web-browsing large language models can access user profiles and infer sociodemographic characteristics from user content. Studies that infer sociodemographics argue that their models could assist costly sociodemographic surveys by offering real-time, cost-effective monitoring. However, those studies also critically discuss the privacy implications of behavioral data such as card data. Indeed, De Montjoye et al. [16] show that by only knowing the place and time of four transactions of a payment card, 90% of card holders can be uniquely identified out of a dataset of 1.1 million cardholders. Card data encode highly private spatiotemporal information about consumers and, therefore, pose privacy issues. A recent study on fraud detection however suggests that the privacy risks associated with sharing sensitive payment data can be mitigated through federated learning, which enables institutions to exchange trained models rather than raw data [5].

In this paper, we propose an approach to enrich card data with sociodemographic information, while complying with the privacy rights of cardholders.

Table 1. Summary statistics of the payment diaries of the Swiss payment surveys (2017 and 2020) and the card dataset.

Data Set	$\#_{participants}$ $\#_{cards}$	$\#_{payments}$	$mean_{value}$	sd_{value}	min_{value}	$Q2_{value}$	$median_{value}$	$Q4_{value}$	max_{value}
2017 Survey	1'966	22'520	39.63	166.28	0.10	7.50	16.10	38.90	12'050
2020 Survey	2'098	21'781	50.00	183.26	0.05	7.70	19.00	47.05	8'000
2020 Card Data	10'844'966	111'145'174	62.00	526.79	0.01	9.40	22.85	54.92	4'000'000



(a) Number of payments per participant

(b) Value per payments in CHF

Fig. 1. Histogram of number and values of payments in the 2017 and 2020 Swiss payment surveys.

3 Data

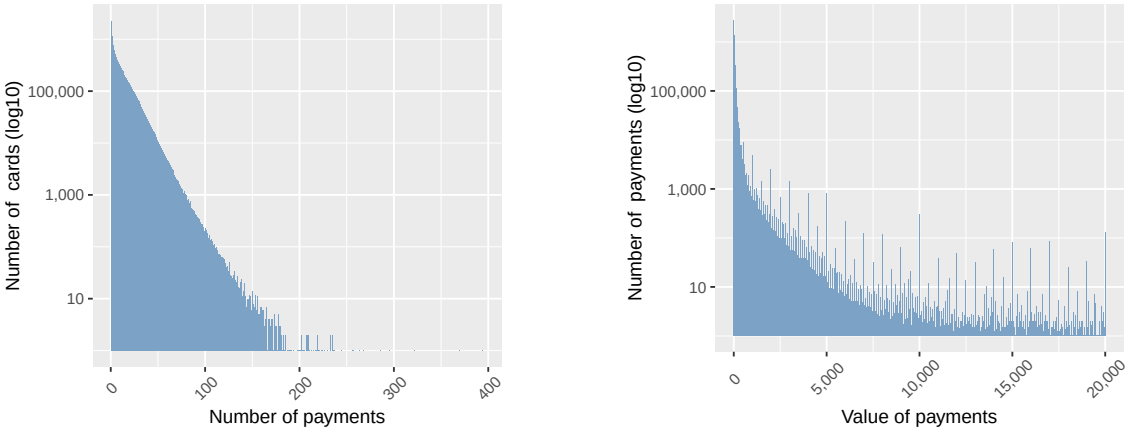
This paper uses several datasets to analyze the payment behavior of sociodemographic groups. The first dataset consists of two payment surveys conducted by the Swiss National Bank. The second dataset consists of anonymized single transaction card data. Finally, publicly available sociodemographic data from the website of the Swiss Federal Statistical Office (FSO) were used to validate the sociodemographic estimates.

3.1 Payment data

In 2017 and 2020, the Swiss National Bank conducted two representative payment surveys that document the payment behavior of the Swiss population [42, 43]. The payment surveys comprise two sections, of which the first consists of an interview with the survey participants, and the second consists of payment diaries, which were filled out by the study participants over a week.

For seven consecutive days, approximately 4,000 random participants from the two studies recorded the value, day, payment purpose, and payment instrument in a payment diary (see Table 1). The payment diaries were self-administered by the study participants over a week. The resulting datasets span 15 weeks in 2017 (08/08/2017 to 25/11/2017, see Appendix, Fig. 9) and 12 weeks in 2020 (18/08/2020 to 16/11/2020, see Appendix, Fig. 9). Fig. 1 shows the payment survey distribution of the number of payments per participant and the values per payment.

The second data source consists of one month of card data collected and provided by PostFinance Ltd. (PostFinance) and Worldline Switzerland Ltd. (Worldline). The two private companies provide payment services to merchants in



(a) Number of payments per card ID

(b) Value per payments in CHF

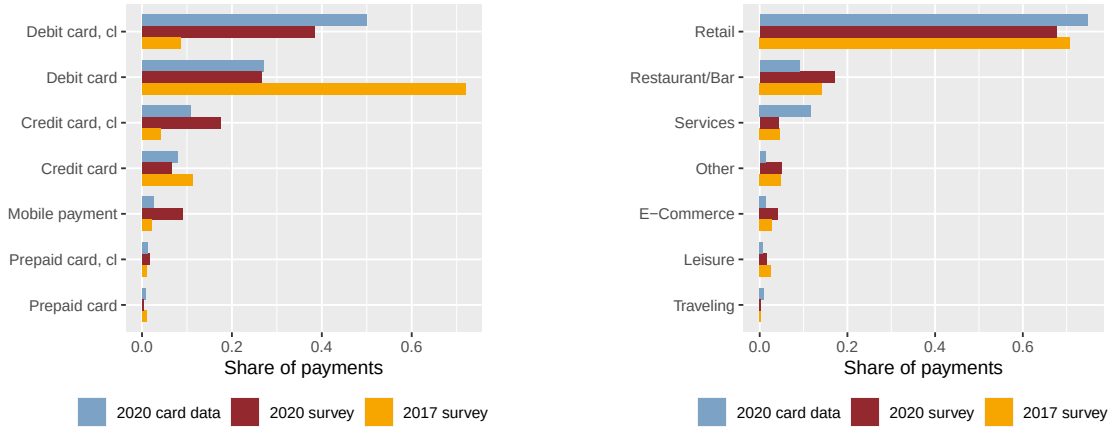
Fig. 2. Histogram of number and values of payments in the card data.

Switzerland. Jointly, the companies process approximately 90% of Swiss card and mobile payments. The two payment providers supply an anonymized and curated card dataset to the Swiss National Bank. In contrast to payment surveys, only electronic payment instruments, which include payment cards such as debit, credit and prepaid cards and mobile payment applications, are recorded. The dataset includes variables such as the payment value, the payment instrument used, a unique and anonymous card ID, the exact time and day, a postal code and a high-level sector categorization of the merchant (two-digit NOGA code as defined in [21]), among other variables. The card dataset is anonymized; therefore, information about individuals are not provided.

From this card dataset, four weeks of payment data were sampled spanning the month of October 2020 (28/09/2020 to 25/10/2020; see Appendix, Fig. 10). During those four weeks, 111 million payments were executed by 10.8 million different payment cards (see Table 1). The distributions of the values of the card data are comparable to the distributions from the surveys. The values in the card data are higher than those in the payment survey data (see Table 1), which can be explained by the fact that low-value cash payments are not present in the card data. The distribution of payments per card, which is the aggregation level that corresponds most closely to a single participant in survey data, can be seen in Fig. 2. On average, there are 4 payments per card per week, which is significantly less than that in the payment surveys. Similarly, as in the survey data, we can observe a highly right-skewed distribution of payments per card. Concerning the distribution of values per payment, we can also observe a highly right-skewed distribution (see Fig. 2 b).

Fig. 3 shows a comparison of payment instrument usage and payment purposes between the 2020 card data and the two payment surveys. To render the different datasets comparable, only payment instruments and payment purposes, which both occur in the card data and the surveys, were considered. This notably excludes cash as a payment instrument, which amounted to 45% in 2020 and 71% in 2017 of all payments.

Concerning the usage of electronic payment instruments, the contactless debit card is by far the most commonly used payment instrument in the 2020 card data and payment survey data, whereas the prepaid card is the least commonly used payment instrument (see Fig. 3 a). While the 2020 card data and survey data generally align in terms of payment instrument usage, there appears to be a shift in payment instrument preferences between 2017 and 2020.



(a) Share of payments per payment instrument (cl = contactless)

(b) Share of payments per payment purpose

Fig. 3. Share payments per payment instrument and purpose as reported in the payment surveys and the card dataset.

To establish consistent payment purposes across both card data and survey data, a mapping between the NOGA codes and the survey-defined payment purposes was created (see Appendix, Table 5). In both datasets, most payments occur in retail stores (see Fig. 3 b). Restaurants and services are also popular in both datasets, although people tend to pay for more services in the card dataset.

3.2 Demographic data

To set payment behavior in relation to sociodemographic variables, two different datasets were used. The first dataset of sociodemographic variables was collected during the personal interview phase of the two payment surveys. These variables include age, sex, marital status, labor force status, income, wealth, immigration status, and residence location. The second sociodemographic dataset consists of national distributions of the previously detailed variables sourced from the Swiss Federal Statistical Office (FSO). The sociodemographic data provided by the FSO are aggregated at the postal code level and include age groups, marital status, and immigration status.

The available demographic information was used to define sociodemographic groups. Considering age two demographic groups were compiled, consisting of people younger than 25 years (Swiss legal definition of youth unemployment) and people aged 65 and above (Swiss legal retirement age). A financial index ranging from 1 to 6 was derived from the participant's income and wealth, as detailed in the payment surveys. Participants with a financial index below or equal to 1.5 belong to a low financial index group (13.3% of participants), whereas people with a financial index of 4.5 and above belong to a high financial index group (14% of participants). Furthermore, sociodemographic variables such as education, sex, marital status, labor force status, immigration status and the residence location of the participants were included in the analysis. All sociodemographic variables, their descriptions, and data shares are detailed in Table 2.

By comparing the sociodemographic distributions of the 2020 payment survey with the 2020 FSO statistics, the representativeness of the payment survey can be assessed. The sociodemographics from the payment studies match

Table 2. Summary table of sociodemographics, the associated binary variables, data shares in the 2017 and 2020 payment surveys, the official 2020 FSO census shares and a short description of each sociodemographic group.

Socio-demographics	Resulting groups	Share 2017 survey	Share 2020 survey	Share 2020 FSO	Description
Age groups	Below 25 (exclusive)	12.7%	13.5%	10.1%	
	Above 65 (inclusive)	17.4%	26.3%	20.6%	The age groups correspond to the Swiss definition of youth unemployment and the legal retirement age in Switzerland. Below 25 does not include people aged 25, while above 65 includes people aged 65. The FSO share is publicly available at [25].
Labour force status	Student	8.28%	11.1%	10.4%	Labour force status derived from the FSO definitions. The FSO share is publicly available at [24].
	Employed	59.3%	57.9%	61.2%	
	Retiree	18.6%	27.2%	22.7%	
Sex	Female	51.6%	50.1%	49.6%	Sex of participants defined as female or male. The following analysis uses female as socio-demographic variable. The FSO shares are publicly available at [25].
	Male	48.4%	49.9%	50.4%	
Marital status	Married	42.7%	52.0%	43.5%	If the participant is married or not. The FSO share is publicly available at [25].
Financial index	Low financial index ($x \leq 1.5$)	25.6%	13.3%	-	The financial index is derived from participant's income and wealth. A low financial index corresponds to low income and wealth and a high index to high income and wealth. The index is normed to a scale from 1 to 6 where personal and household as well as income and wealth are equally weighted. No comparable FSO data are available.
	High financial index ($4.5 \leq x$)	7.22%	14%	-	
Education	Primary	13.4%	9.13%	24.5%	The highest completed education of participants. Educations are categorised in primary (including compulsory education, etc.), secondary (including apprenticeship, vocational school, academic matur, etc.), and tertiary (including professional or technical school, university, etc.) education. The definition of the categories as well as the official FSO shares are publicly available at [22]
	Secondary	65.7%	48.0%	38.3 %	
	Tertiary	20.8%	42.9%	37.2%	
Immigration status	Swiss citizenship	84.8%	85.7%	82.9%	If the participant possesses Swiss citizenship or not. The FSO share is publicly available at the [25] .
Residence location	Urban	70.8%	61.1%	62.8%	If the participant lives in an urban, suburban or rural community. Communities are defined as urban, suburban, or rural using the the FSO geographical typology. The FSO shares and the definitions of the categories are publicly available at [23]. .
	Suburban	15.7%	22.1%	21.8%	
	Rural	13.5%	16.8%	15.5%	

the official shares to different degrees (see Table 2). The highest absolute difference amounts to 15% (primary school-educated people). The average absolute difference between the FSO shares and both surveys amounts to approximately 4.5%. The sociodemographic groups and resulting variables are in a binary format (only two groups per variable).

4 Methodology

First, this paper applies feature engineering to construct several feature sets. Second, we employ statistical methods to analyze and identify payment behavior across sociodemographic groups. Third, we present the machine learning methods used to infer sociodemographic variables. Finally, we describe the inference pipeline designed to enrich anonymized card data with sociodemographic information.

4.1 Feature extraction

To analyze the payment behavior of sociodemographic groups, different sets of payment features were extracted from the payment surveys. To obtain the payment behavioral features, the recorded payments were summarized for each participant according to common metrics and three dimensions (time, purpose, and payment instrument). Table 6 in the Appendix summarizes the feature extraction per feature set and per feature dimension. In total, three different feature sets were generated: an extensive, simplified, and a card data-emulated feature set. The extensive feature set consists of 103 different variables describing the payment behavior of participants. The simplified feature set consists of 79 variables and was obtained by simplifying several variables. Compared with the extensive feature set, the simplified feature set

contains more readable features. Student's t-tests were used to explore significant differences in the payment behavior of different demographic groups. The card data-emulated feature set was constructed to emulate the anonymized card data and consists of 34 variables. The card data-emulated feature set summarizes the payment behavior per payment instrument. This feature set only contains electronic payment instruments such as debit, credit, prepaid cards, and mobile payment applications. The card data-emulated feature set is used to train models that infer sociodemographic distributions in the card data. The payment data were aggregated according to three dimensions: a time dimension (e.g., the mean value of payments on Tuesday), a purpose dimension (e.g., the number of payments in retail stores), and a payment instrument dimension (e.g., the number of payments with credit cards). For every dimension, the payment distribution per study participant was summarized with simple metrics (e.g., the mean value of payments) following the sensing and social computing literature [46].

4.2 Statistical methods

In the second stage, to assess which payment behavior is most indicative of a demographic group, independent Student's t-tests [32] are performed on the simplified feature set. P values were adjusted with the Bonferroni correction [45]. To account for differing variances per group, t-tests were performed with a Welch correction [17]. Additionally, the mean differences between payment features per demographic group with 95% confidence intervals were calculated. The mean differences are used to assess the noncausal effect size of a demographic variable on a feature. The mean differences are denoted in units of the respective feature (e.g., the value is denoted in CHF). Additionally, the mean differences have been standardized with Cohen's d metric [35].

4.3 Machine learning methods

To assess whether sociodemographic variables can be inferred by payment behavior, an inference task was defined. The inference task was run on the 2020 and 2017 payment survey data via the extensive feature set. All the demographic variables are inferred in a one-vs.-many binary setting. To address class imbalances, naive downsampling was used in training and testing. Given the resampling and binarization of the variables, all inference results can be compared with a 50% baseline accuracy.

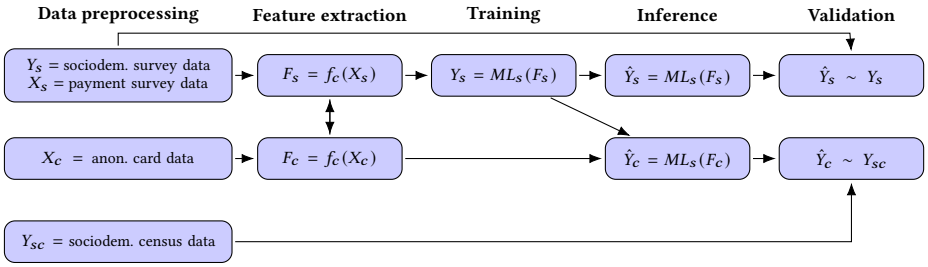
The machine learning methods used in the inference task include regularized logistic regression, random forest, XG Boost, Ada boost and a fully connected multilayer neural network (all methods are referenced in [41]). All the experiments were implemented with python and run with machine learning libraries such as scikit-learn [38] and a python XG Boost implementation [12] and PyTorch [37] for the neural network. Hyperparameter tuning was performed on all methods via exhaustive grid search cross-validation [41]. For the random forest, parameters such as the number of trees, minimum sample split, and maximum depth were optimized. For XG Boost, the number of estimators, the max depth, the penalty value (lambda), and the learning rate were optimized. The neural network consists of four hidden layers (100, 80, 40, and 20 neurons) and uses leaky ReLU as the activation function for the hidden layers and a sigmoid as the activation function for the output layer. The number of epochs, batch size, and learning rate were determined via hyperparameter tuning.

4.4 Inference pipeline

To enrich anonymized card data with sociodemographic information, we use an inference pipeline, which comprises five stages: Data preprocessing, feature extraction, training, inference and validation (see Fig. 4). The pipeline uses all discussed data sources: payment and sociodemographic survey data, anonymized card data, and sociodemographic

census data (see Section 3). The data were preprocessed by removing observations with missing values and ensuring that the payment survey and card data shared the same set of attributes, for instance by mapping NOGA codes to survey-defined payment locations (see Section 3.1). Then, we extracted the card data-emulated feature sets (see Section 4.1) from both the payment survey and the card data. It is important to point out that we used the same feature extraction method (in Fig. 4 referred to as $f_c()$) for payment survey and card data, so to later infer with the same model sociodemographic information from both payment survey and card data. Using the payment features and sociodemographic variables extracted from the payment survey data, we trained models with the machine learning methods described in Section 4.3. In our context, a model refers to a trained machine learning method that can infer sociodemographic information from unseen payment data. The models were then applied to infer sociodemographic variables from both survey and card data. In the final step, the inference results were validated (referred to as $\hat{Y}s \sim Ys$ in Fig. 4). Validation on survey data is straightforward given that it contained sociodemographic variables. Validation followed a 10-fold cross-validation strategy, which provides mean accuracies and their standard errors. Furthermore, additional metrics as recall, precision, and F1 were computed. For the sociodemographic inference on anonymized data, validation was performed by comparing inferred sociodemographic variables at the postal-code level with census data. Assuming that individuals make payments where they live, the correlations between the model-inferred and census distributions indicate how well sociodemographic variables can be inferred from anonymized card data.

The proposed inference pipeline provides a reproducible framework for public institutions and can be regularly retrained on new payment survey data to mitigate sample bias. This is particularly important because our survey was conducted during the COVID-19 pandemic and captures only one week of payment data per participant. Moreover, the data may suffer from selection bias. Although the survey targeted a representative sample of the Swiss population, participants were incentivized with 100 CHF (approximately 110 USD in 2022), which likely increased participation among lower-income individuals, who are indeed overrepresented (see Table 2). Such biases heighten the risk of distributional shifts, where the training data distribution differs from the inference data distribution [29, 30, 36], thereby reducing inference performance. Thus, continuous retraining and fine-tuning with new data are essential for improving robustness and minimizing bias in the inference pipeline.



Variables: Y = sociodemographic variables, X = payment data, F = payment features, $\hat{Y}$ = inferred sociodemographic variables, $f()$ = feature extraction method, $ML()$ = machine learning method,
Subscripts: s = survey data, c = anonymized card data, sc = sociodemographic census data

Fig. 4. Flow diagram of the inference pipeline for enriching anonymized card data with sociodemographic information.

5 Results

In this section, we first show that sociodemographic groups significantly differ in their payment behavior. Then we conduct an analysis to assess how well those characteristics can be inferred from payment survey data. Finally, we

show that card data can be used to infer sociodemographics that align with national distributions, and we explore a potential application of card data enriched with sociodemographics.

5.1 Statistical analysis

This section explores which payment behavior is most indicative of different sociodemographic groups. The payment behavior of sociodemographic groups was modeled with the simplified payment feature set. In Table 3, the top 5 most indicative features per sociodemographic group, as measured by Student's t-tests on the basis of the 2020 data, are shown. Table 7 in the Appendix shows summary statistics per binary demographic group. Table 8 in the Appendix shows the top 5 most indicative features per sociodemographic group on the basis of the 2017 data. The most indicative features of the 2017 and 2020 surveys are similar.

We discuss the most indicative payment behavior of age groups, labor force, sex and marital status, financial index groups and education level. The most indicative features per demographic group are often intuitive and in accordance with the literature.

Considering age groups, participants below 25 years spend, on average, 28 CHF less in retail stores and have 2.3 cash payments less per week than older participants. People above 65 have, on average, 2.5 more cash payments, while using the contactless function less frequently. Furthermore, older participants are less likely to use mobile or e-commerce payments. The observed differences in age-related payment behavior are in accordance with the literature [6, 11, 14].

Considering labor force status, participants who are employed tend to have more payments than people who are not employed. Specifically, employed participants have more contactless, debit, e-commerce and mobile payments. With respect to sex and marital status, we find that women tend to make fewer payments at restaurants, bars, and petrol stations while making more payments for durable products. Compared with unmarried participants, married participants tend to have greater expenditures at retail stores and petrol stations but go less frequently to restaurants and bars.

Considering the financial index of participants, low-financial index participants have very constant and low value payments in restaurants and bars. This behavior may reflect routine visits to cafeterias during lunchtime. Furthermore, participants with a low financial index tend not to use credit cards and contactless payments. In contrast, high-financial index participants often use credit cards, the contactless feature, and e-commerce platforms. Their expenditure in restaurants is, on average, 15 CHF higher. These findings are in accordance with prior research, which suggests that low-income people tend to use cash and that high-income people tend to use credit cards [14, 34].

Finally, the education level of the participants seemed to influence their payment behavior. People with primary education make fewer payments in general while having fewer payments for public services in particular. Furthermore, their mean expenditure with credit cards is 27 CHF lower. People with tertiary education, however, seem to be quite technologically adept. They often pay with the contactless feature, credit cards, mobile payments, and on e-commerce platforms. The results regarding the differences in cash and credit card usage according to education level are supported by other studies [14, 34].

In conclusion, the statistical analysis revealed that the payment behavior of sociodemographic groups differed significantly. The use of payment instruments, especially cash and cards, depends strongly on age, educational level, and financial status.

Table 3. Top 5 most indicative features as measured by Student's t tests for each sociodemographic group in the 2020 survey.

	Feature	t_stat	Diff	[95% CI Diff]	CD		Feature	t_stat	Diff	[95% CI]	CD
Below 25	$\overline{value}$ Retail	16.36****	-28.06	[-31.42, -24.69]	-0.55	Above 65	#Cash	11.13****	2.48	[2.04, 2.92]	0.60
	#Cash	11.08****	-2.25	[-2.65, -1.85]	-0.54		#contactless	9.6****	-1.58	[-1.9, -1.25]	-0.43
	#Payments	10.86****	-3.13	[-3.7, -2.56]	-0.59		#Debit Card	8.35****	-1.37	[-1.69, -1.05]	-0.38
	#Household services	10.25****	-0.07	[-0.09, -0.06]	-0.26		#E-Commerce	7.41****	-0.25	[-0.31, -0.18]	-0.29
	#Retail	10.19****	-1.83	[-2.18, -1.48]	-0.58		#Mobile/Online	7.17****	-0.46	[-0.58, -0.33]	-0.29
Student	$\overline{value}$ Retail	15.03****	-27.27	[-30.83, -23.7]	-0.53	Employed	#contactless	9.62****	1.50	[1.2, 1.81]	0.41
	$\overline{value}$ Petrol stations	12.29****	-12.31	[-14.28, -10.35]	-0.48		#Debit Card	7.47****	1.15	[0.85, 1.45]	0.32
	#Payments	11.89****	-3.46	[-4.03, -2.89]	-0.65		#E-Commerce	6.03****	0.22	[0.15, 0.29]	0.25
	#Retail	11.54****	-2.08	[-2.43, -1.72]	-0.66		#Mobile/Online	5.63****	0.37	[0.24, 0.5]	0.23
	#Cash	10.22****	-2.19	[-2.61, -1.77]	-0.52		#Payments	5.46****	1.29	[0.83, 1.76]	0.24
Female	#Restaurant/bar	8.9****	-0.95	[-1.16, -0.74]	-0.39	Married	$\overline{value}$ Petrol stations	6.19****	6.91	[4.72, 9.1]	0.27
	#Petrol stations	6.37****	-0.25	[-0.32, -0.17]	-0.28		$\overline{value}$ Retail	5.48****	12.49	[8.02, 16.96]	0.24
	$\overline{value}$ Restaurant/bar	5.45****	-8.59	[-11.68, -5.5]	-0.24		$\sigma(\overline{value})$ Credit card	5.06****	13.77	[8.43, 19.1]	0.22
	#Durable products	5.08****	0.25	[0.16, 0.35]	0.22		$\overline{value}$	5.02****	14.55	[8.86, 20.24]	0.22
	$\overline{value}$ Petrol stations	4.43***	-5.02	[-7.24, -2.8]	-0.19		#Restaurant/bar	4.89****	-0.53	[-0.75, -0.32]	-0.22
Low fin	$\sigma(\overline{value})$ Restaurant/bar	6.46****	-7.60	[-9.91, -5.29]	-0.23	High fin	#Credit card	6.06****	1.27	[0.86, 1.68]	0.51
	#Credit card	6.04****	-0.79	[-1.05, -0.53]	-0.31		$\overline{value}$ Restaurant/bar	5.13****	15.39	[9.48, 21.29]	0.42
	$\overline{value}$ Credit card	6.04****	-20.91	[-27.7, -14.12]	-0.24		#contactless	4.82***	1.30	[0.77, 1.83]	0.35
	$\overline{value}$ Restaurant/bar	5.85****	-9.81	[-13.1, -6.52]	-0.27		#E-Commerce	4.62***	0.35	[0.2, 0.5]	0.40
	#Mobile/Online	5.74****	-0.39	[-0.52, -0.25]	-0.24		$\overline{value}$ Credit card	4.29**	25.57	[13.84, 37.3]	0.29
Primary edu	$\overline{value}$ Credit card	9.42****	-26.84	[-32.44, -21.25]	-0.31	Tertiary edu	#contactless	7.16****	1.21	[0.88, 1.54]	0.32
	#Payments	8.25****	-2.84	[-3.51, -2.16]	-0.53		#E-Commerce	6.5****	0.26	[0.18, 0.34]	0.30
	#Public services	7.15****	-0.06	[-0.08, -0.04]	-0.22		#Credit card	6.31****	0.73	[0.51, 0.96]	0.29
	$\sigma(\overline{value})$	6.58****	-38.51	[-50, -27.01]	-0.24		#Mobile/Online	6.15****	0.46	[0.32, 0.61]	0.29
	$\overline{value}$ Online banking	6.36****	-45.52	[-59.54, -31.49]	-0.17		#Payments	5.89****	1.41	[0.94, 1.88]	0.26

5.2 Inferring sociodemographic groups in the survey data

The following section investigates whether sociodemographic groups can be inferred by their payment behavior, as detailed in the payment survey. First, different models are trained and tested with the extensive feature set on the 2020 survey data. Second, the accuracies obtained per sociodemographic group are discussed. Third, the results of the 2020 survey are compared with the results of the 2017 survey.

Table 4. Results per sociodemographic group for logistic regression, Random Forest, XG Boost, Ada Boost and multilayered neural network.

Demographic	$\#_{samples}$	**Logit**		**Random Forest**		**XG Boost**		**Ada Boost**		**Neural Net**	
		$\overline{A}$	(A_σ)	$\overline{A}$	(A_σ)	$\overline{A}$	(A_σ)	$\overline{A}$	(A_σ)	$\overline{A}$	(A_σ)
Students	(468)	61.03	(2.63)	83.11	(1.31)	82.90	(1.5)	**83.51**	(1.33)	79.04	(1.39)
Below 25	(566)	62.84	(1.57)	**82.00**	(1.84)	81.09	(1.76)	80.74	(1.47)	77.94	(2.31)
Above 65	(1100)	61.64	(1.41)	71.00	(1.14)	**71.64**	(1.39)	69.91	(1.08)	65.82	(1.46)
In retirement	(1136)	63.74	(1.06)	70.96	(1)	70.60	(0.9)	**71.39**	(0.75)	64.27	(1.7)
Primary edu.	(376)	65.95	(1.88)	**70.42**	(2.47)	67.73	(3.63)	67.79	(2.19)	63.29	(1.88)
High fin.	(590)	62.53	(2.15)	**67.30**	(1.73)	65.25	(1.02)	64.91	(1.36)	59.79	(1.96)
Married	(2016)	61.81	(0.91)	**65.92**	(0.79)	65.28	(1.06)	65.13	(1.05)	62.81	(0.89)
Employed	(1764)	58.51	(1.07)	**64.92**	(1.22)	64.52	(1.11)	64.12	(1.01)	59.75	(1.11)
Female	(2094)	59.17	(0.77)	63.70	(1.01)	63.89	(0.88)	**64.90**	(0.77)	58.69	(0.52)
Low fin.	(558)	60.07	(1.53)	**64.37**	(2.35)	61.81	(1.65)	60.24	(2.23)	58.44	(1.8)
Tertiary edu.	(1796)	58.74	(1.04)	**61.92**	(0.89)	60.30	(0.78)	60.80	(0.71)	56.91	(0.81)
Countryside	(704)	53.70	(1.1)	**58.67**	(2.05)	57.52	(1.58)	56.11	(1.59)	52.59	(1.98)
City	(1628)	**56.26**	(1.41)	55.28	(0.92)	54.97	(1.22)	52.22	(0.83)	52.15	(0.95)
Swiss citizenship	(598)	49.37	(1.9)	55.53	(1.85)	**56.21**	(1.82)	54.54	(1.81)	51.50	(1.38)

First, Table 4 shows the results for the five machine learning models applied to the full set of 15 sociodemographic variables and the 2020 extensive feature set. A comparison of the different machine learning models reveals that the

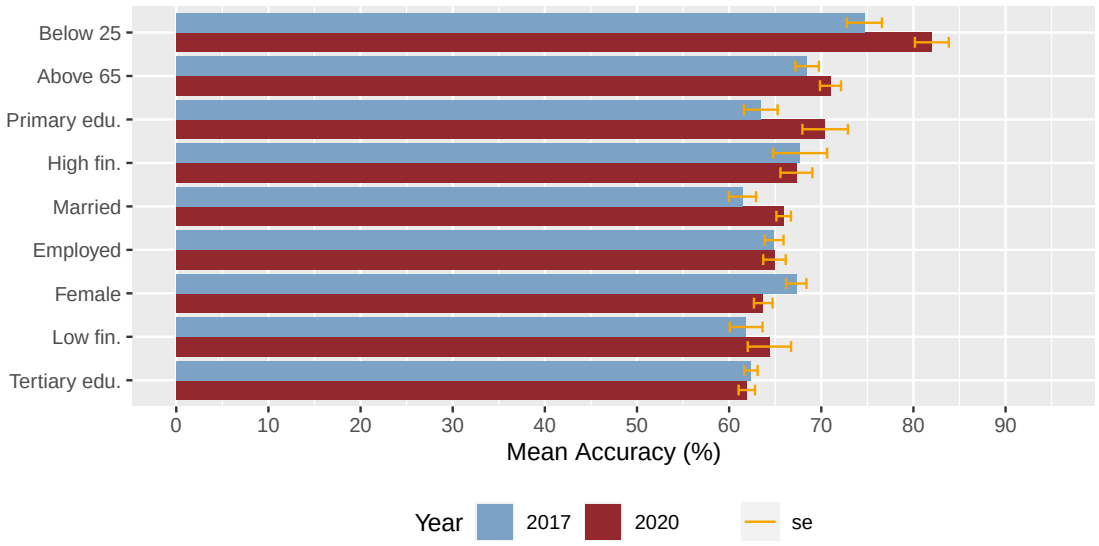


Fig. 5. Sociodemographic inference results on the basis of the 2017 and 2020 payment survey data.

ensemble models (random forest, XG Boost, AdaBoost) outperform the neural network and the logistic regression. The three ensemble models produce very similar results. However, the random forest outperforms the two other models by a small margin. The random forest especially outperforms the other two models in the case of inferences with a limited sample size. The inference class imbalances, as measured by the precision and recall metrics, of the ensemble models are negligible (see Appendix, Table 9). For all subsequent experiments, a random forest is used.

Second, as seen in Table 4, the random forest infers binary sociodemographic groups with accuracies ranging between above 80% to below 60%. Some demographic groups that are especially well identifiable by their payment behavior are age groups and closely related labor market groups (accuracies ranging between 71% and 83%). However, given the strong overlap of those groups (people above 65 years tend to be in retirement), in the subsequent experiments, only the results of the age groups are shown. Participants with primary education can be inferred with an accuracy above 70%. Other variables such as marital status, sex, employment status, and the financial index can be inferred with accuracies ranging between 60% and 70%. Finally, there are variables with accuracies below 60% and high standard errors. Those variables include the residence location (city and countryside) and the participants' nationality. Those sociodemographic groups seem to be almost not discernible given their payment behavior. This finding indicates that there are no specificities in the payment behavior of those groups or that those sociodemographic groups are ill defined. All variables with an accuracy below 60% were not considered in subsequent experiments.

Third, when we compare the results of the selected sociodemographic groups on the basis of the data from 2017 and 2020, we can observe some differences (see Fig. 5). First, the 2020 data tend to yield higher accuracies ($\Delta\overline{A}2.2\%$). For example, the model infers the participants below 25 group and the primary educated group with 7% accuracy better in 2020 than in 2017. However, for most variables, the difference is less than 5%. Finally, the order of accuracies remained very similar in 2017 and 2020. This means that sociodemographic groups that can be accurately inferred in 2017 can also be accurately inferred in 2020.

In conclusion, we show that a wide variety of sociodemographic groups can be identified accurately by their payment behavior. Some sociodemographic groups, such as younger people (below 25) or older people (above 65), have very discernible payment behavior. Other sociodemographic variables, such as the nationality or the residency location, cannot be identified by payment behavior. Group-specific payment behavior persists over time (see Fig. 11) and is more accentuated in 2020 than in 2017.

5.3 Inferring sociodemographic groups in the card data

The following section investigates whether the learned relationship between payment behavior and sociodemographic groups can be used to infer sociodemographic groups in anonymized card data. We proceed in three stages. First, models are trained and tested on the card data-emulated feature set from the 2020 survey to estimate accuracies in an anonymized setting. Second, trained models are employed on the 2020 card data to estimate sociodemographic distributions and are validated by comparing the inferred distributions with postal code census data. Third, time-dependent age distributions in the card data have been estimated as a possible monitoring use case.

In the first stage of inference, the model is trained on the card data-emulated feature set, which mimics the features that can be created with the anonymized card data. In contrast to the survey data, sociodemographics are inferred here on the level of cards. Thus, the models have no information about payment instrument choices. In Fig. 6, the resulting accuracies of the card data-emulated and extensive feature sets are compared on the basis of the 2020 survey data. It is apparent that there is a large decrease of 7.1% in mean accuracy across all sociodemographic groups. The decrease in accuracy reinforces the sentiment that payment instrument choices contain important sociodemographic information. Nevertheless, 4 variables have accuracies above 60%. The best identifiable sociodemographic groups are age groups. The age group below 25 reaches an accuracy of 72%, and the age group above 65 reaches an accuracy of 65%, with both variables having standard errors of approximately 1.4%. Married people were identified with an accuracy of 62.4% and low standard errors. For the following experiments, only the married group and the age groups below 25, and above 65 were considered.

In the second stage of inference, the sociodemographic variables below 25, above 65, and married were inferred from anonymized card data. To validate the estimated sociodemographic variables, the estimated distributions were compared with official census data published by the Federal Statistical Office [25]. To aggregate the inferred sociodemographics, it was assumed that the cardholders live in the postal code where they make the most payments. The estimated sociodemographic distribution of each postal code was subsequently compared with the census data. To obtain robust estimations, only postal codes with more than 5,000 inhabitants were considered, resulting in 402 valid postal codes. The resulting correlations between the census data and the inferred distributions of the card data are shown in Fig. 7. The inferred shares of the age group below 25 do not match the official shares and produce a negligible negative correlation. The age group above 65 reaches a moderate correlation of $R = 0.39$. Finally, the inferred shares of married people match the official shares well, producing a strong correlation of $R = 0.66$. The negative correlation concerning the age group below 25 may either indicate that the model does not identify young people reliably or that the assumption for the spatial mapping of inferred shares and census shares does not hold for young people. The assumption that people live where they make the most payments is disputable and may breakdown with very mobile population groups. For example, young people may live with their parents in suburbs while commuting every day to university or their workplace in the city, where they have lunch and do their shopping.

In the third stage of inference, we explore the informative value of the in card data inferred sociodemographics. As a use case we calculate the likelihood of payments being made by age groups during different times of the day (see Fig. 8).

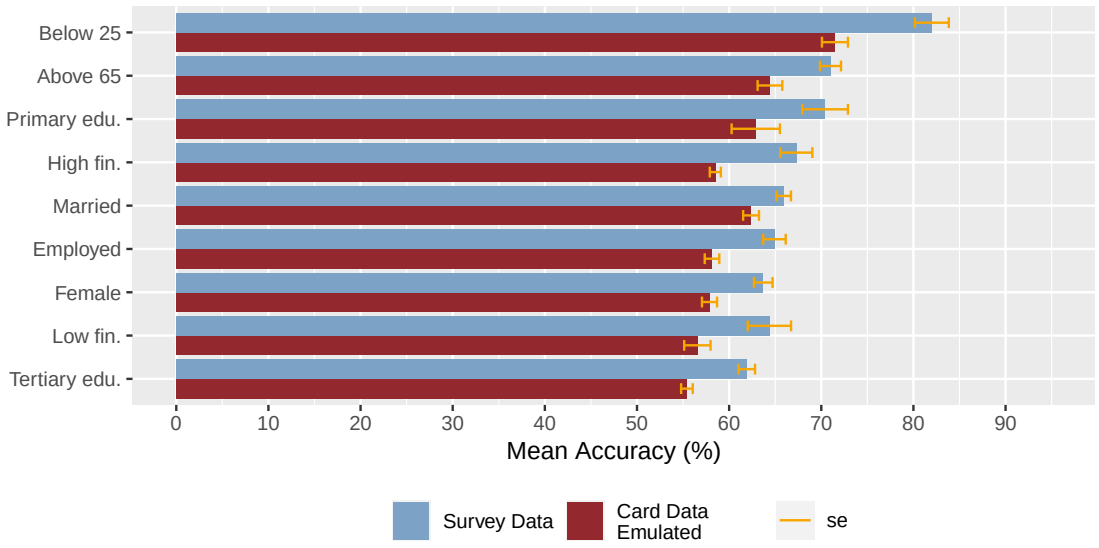
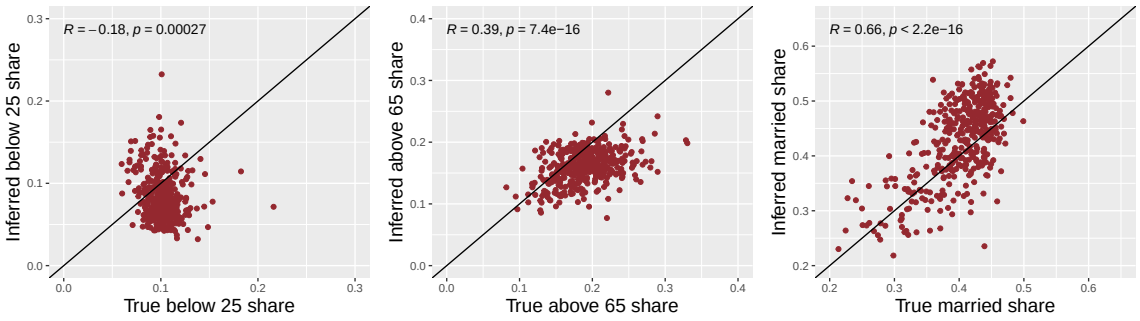


Fig. 6. Sociodemographic inference results on the basis of the card-emulated feature set compared with the extensive feature set.



(a) Correlation of inferred and true share of the below 25 age group

(b) Correlation of inferred and true share of the above 65 age group

(c) Correlation of inferred and true share of married people

Fig. 7. Correlation analysis between inferred sociodemographic shares and census sociodemographic shares. The R correlation measurement and the p value (different from zero) are reported in the upper left corner.

This is an interesting monitoring use case, given that the models never had information on the time of payments. To determine the normalized likelihood of an age group making a payment during a specific hour, the payment probability per hour of each age group was estimated and normalized via z-normalization. A negative normalized likelihood indicates that an age group is less likely to make a payment during a given hour. A positive likelihood indicates that the age group is more likely to make a payment. The resulting likelihoods of the two age groups mirror each other. This is partially mechanically driven given that a low likelihood of one age group leads to a high likelihood of the opposed age group. However, it is interesting to examine when the models suggest that one age group (or the other) is active. Looking at the normalized likelihoods during weekdays (see Fig. 8), we can see that the model estimates that people

(a) During weekdays (b) During weekends

Fig. 8. Normalized inferred likelihood of payments being made by below-25 or above-65 age groups given the hour of the day.

below 25 are rather active during the morning, lunchtime, and the evening, whereas people aged 65 and above are doing their shopping before and after lunch. During the weekend, payments between 7:00 and 18:00 are estimated to be from older people, whereas payments after 20:00 and into the night and early morning are estimated to be from younger people. The models intuitively suggest that the payment behavior of different age groups depends on the time of day and differs between weekdays and weekends.

In conclusion, the experiments using survey data (Fig. 6) show that it is possible to identify sociodemographic variables in card data. Furthermore, it is possible to estimate postal code-level sociodemographic distributions on the basis of card data for some variables (see Fig. 7). Finally, the models intuitively suggest that different age groups may have different temporal payment patterns throughout the day (see Fig. 8).

6 Discussion

This analysis shows that sociodemographic groups exhibit fundamentally different payment behaviors concerning the choice of payment instrument and payment behavioral metrics. It is possible to identify and infer sociodemographic groups in payment survey data. Estimating sociodemographic shares in anonymized card data is more difficult, but promising results have been obtained. To fairly assess this study, it is necessary to contextualize the presented results concerning policy implications, privacy considerations, data limitations, and reliability.

6.1 Policy implications

The policy implications of this study are twofold. First, different population groups show different preferences for payment instruments. Second, the proposed inference pipeline allows sociodemographic characteristics to be inferred from payment behavior, enabling public institutions to enrich card data with sociodemographic information.

With respect to the first point, our statistical analysis (Section 5.1) shows that the use of different payment instruments depends on sociodemographic characteristics such as age, sex, educational level, income, and marital status. The t-tests in Table 3 indicate, for example, that university-educated and high-income groups are more familiar with new payment

technologies such as contactless and mobile payments. By contrast, people with lower incomes and those above 65 tend to rely on cash (see Table 3). This suggests that the ongoing trend of declining cash usage and increasing usage of electronic payment instruments [4, 13, 44] does not apply equally across all parts of society. Thus, to ensure financial inclusion, it remains important to maintain access to payment instruments with declining usage, particularly cash.

Regarding the second point, our inference results with payment survey data (see Section 5.2) show that some sociodemographic groups can be identified from payment behavior. Moreover, our inference results with card data (see Section 5.3) indicate, that it is possible to estimate the distribution of certain sociodemographic groups in anonymized card data. This enables ongoing monitoring of payment behavior across sociodemographic groups using card data.

The inference pipeline described in Section 4.4 enables public institutions to use card data to monitor sociodemographic groups alongside traditional payment surveys. In this way, the methodology can serve as a cost-effective complement to surveys by providing more timely insights between survey waves. This might not only result in lower costs associated with surveys but also reduce the burden on participants and allow updates to be made more frequently. Such continuous monitoring may be useful for tracking payment trends or reactions to economic shocks across sociodemographic groups, while still ensuring that the privacy rights of cardholders are respected.

6.2 Privacy considerations

Both the data used in the study and the investigated research questions can be discussed concerning possible privacy implications. With respect to the data, the payment surveys were conducted following a strict privacy protocol [42, 43]. No individuals can be identified in the survey data. The used card data were anonymized, and data protection measurements were implemented to render the identification of individuals impossible. This includes limiting location information of payments to coarse geographic information and indicating broad sector information of merchants. The data storage method complies with data protection regulations and is accessible only from devices within the Swiss National Bank network. With respect to the research questions, this study only considers the aggregate of a given sociodemographic group. This study does not identify the sociodemographics of individuals but identifies the payment behavior of sociodemographic groups.

6.3 Data limitations and reliability of the results

Concerning the data, two limitations must be noted. First, the data represent a small and biased time frame. Second, two datasets with different data definitions were used jointly. To fairly assess the reliability of the results, we must consider first that sociodemographics can be ill-defined and second that some assumptions to estimate sociodemographic distributions are disputable.

First, the payment diary data cover only a short and potentially biased time period. As discussed in Section 4.4, participants recorded their payments for just one week, were incentivized with 100 CHF, and the study took place during the COVID-19 pandemic. These factors may have increased the likelihood of sample bias.

Second, the payment surveys and the card data present differing definitions of payment instruments and payment purposes. For example, credit cards used with a mobile wallet (e.g., Apple Pay) are defined as mobile payments in the payment surveys and as credit card payments in the card data. This may be problematic given that the inference models are trained on the payment surveys and then employed on the card data. Such differences in data specifications may lead to distributional shifts, which may bias the inference [29, 30, 36].

Considering the reliability of the results, two points are worth discussing. First, we must discuss the definitions of sociodemographic groups. While some sociodemographic groups can be inferred with accuracies exceeding 80%, others

can be inferred only moderately well. This is not surprising given that sociodemographic groups do not behave uniformly or may be ill-defined [39]. Some participants below 25, for example, already have a stable job with a high-income, which is not a common trait for this age group. The obtained results describe the average representative of a sociodemographic group and should certainly not be generalized to all members of the group.

Second, this study estimates sociodemographic distributions with anonymized card data, which match official census data with correlations up to $R = 0.66$ but also negative correlations. The negative correlations between the official census data and the obtained estimations may suggest that the models produce unreliable results with card data. However, the assumption that people live where they pay, which allows the mapping of sociodemographic census shares to card data shares, may break down with mobile population groups. Future work could focus on establishing other validation strategies that test whether estimated sociodemographic distributions in unlabeled data are reliable.

7 Conclusion

This study shows that sociodemographic groups exhibit different payment behaviors and proposes a methodology to enrich card data with sociodemographic information.

This study discusses differences in payment behavior across sociodemographic groups. Based on more than 4,000 payment diaries from a 2017 and 2020 payment surveys, we identify distinct payment behavioral patterns per sociodemographic group. Older and low-income groups rely more heavily on cash, whereas younger, higher-income, and tertiary-educated groups tend to use more digital payment methods such as mobile, contactless, and e-commerce payments. These findings indicate that the current decline in cash usage is not uniform across society. This highlights the need for public institutions to maintain access to cash infrastructure in order to ensure financial inclusion across all parts of society.

This study proposes a methodology that enables public institutions to monitor the payment behavior of sociodemographic groups using card data while safeguarding privacy. Based on payment and sociodemographic survey data, we show that sociodemographic groups can be inferred from their payment behavior, with accuracies ranging from 55% to 83%. Building on this result, we propose an inference pipeline that allows public institutions to enrich anonymized card data with sociodemographic information. The pipeline applies machine learning models trained on survey data to estimate sociodemographic distributions in anonymized card data. It includes a validation mechanism that compares these inferred distributions with national census data. By adopting this methodology, public institutions can complement regular surveys with timely and cost-effective information on sociodemographic payment behavior from anonymized card data.

Future work could extend this study in several ways. First, since our analysis relies on temporally limited survey data from Switzerland collected during the COVID-19 pandemic, the inference pipeline should be validated using data from other periods and countries. Ideally, this would involve longitudinal survey data tracking the same individuals over time, enabling continuous evaluation of the pipeline's reliability and model retraining. Second, public institutions could benefit from integrating the pipeline into real-time payment data acquisition processes. Investigating how to adapt the approach for real-time sociodemographic inference would raise important questions regarding validation strategies and the mitigation of biases from distributional shifts. Third, our current validation relies on comparisons with national census data, which assumes that people make payments where they live. As discussed in the paper, this assumption might not hold for certain sociodemographic groups. Future research should therefore develop more robust validation strategies for sociodemographic inference in anonymized card data.

References

[1] Knut Are Aastveit, Tuva Marie Fastbø, Eleonora Granziera, Kenneth Sæterhagen Paulsen, and Kjersti Næss Torstensen. 2020. Nowcasting norwegian household consumption with debit card transaction data. (2020).

[2] Aditya Aladangady, Shifrah Aron-Dine, Wendy Dunn, Laura Feiveson, Paul Lengermann, and Claudia Sahm. 2019. *From transactions data to economic statistics: constructing real-time, high-frequency, geographic measures of consumer spending.* Technical Report. National Bureau of Economic Research.

[3] Meysam Alizadeh, Fabrizio Gilardi, Zeynab Samei, and Mohsen Mosleh. 2025. Web-Browsing LLMs Can Access Social Media Profiles and Infer User Demographics. *arXiv preprint arXiv:2507.12372* (2025).

[4] Guerino Ardizzi, Andrea Nobili, and Giorgia Rocco. 2020. A game changer in payment habits: evidence from daily data during a pandemic. *Bank of Italy Occasional Paper* 591 (2020).

[5] Tomisin Awosika, Raj Mani Shukla, and Bernardi Pranggono. 2024. Transparency and privacy: the role of explainable ai and federated learning in financial fraud detection. *IEEE access* 12 (2024), 64551–64560.

[6] John Bagnall, David Bounie, Kim P Huynh, Anneke Kosse, Tobias Schmidt, Scott D Schuh, and Helmut Stix. 2014. Consumer cash usage: A cross-country comparison with payment diary survey data. (2014).

[7] Ali B Barlas, Seda Guler Mert, Berk Orkun Isa, Alvaro Ortiz, Tomasa Rodrigo, Baris Soybilgen, and Ege Yazgan. 2021. Big Data Information and Nowcasting: Consumption and Investment from Bank Transactions in Turkey. *arXiv preprint arXiv:2107.03299* (2021).

[8] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd international conference on World Wide Web*. 131–140.

[9] Juliana B Bolzani. 2022. Leading the way in payments: how central banks are using innovation to promote financial inclusion and reshape competition. *JL & Com.* 41 (2022), 103.

[10] Martin Brown, Laura Felber, and Christoph Meyer. 2025. Consumer adoption and use of payment technology: Convenience benefits vs. security concerns. *Working Paper, Swiss National Bank, Study Center Gerzensee* No. 25.01 (2025).

[11] Martin Brown, Nicole Hentschel, Hannes Mettler, and Helmut Stix. 2020. Financial innovation, payment choice and cash demand–causal evidence from the staggered introduction of contactless debit cards. (2020).

[12] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[13] Gabriele Coletti, Alberto Di Iorio, Emanuele Pimpini, and Giorgia Rocco. 2022. Report on the payment attitudes of consumers in Italy: results from ECB surveys. *Bank of Italy Markets, Infrastructures, Payment Systems Working Paper* 21 (2022).

[14] Sean Connolly and Joanna Stavins. 2015. Payment Instrument Adoption and Use in the United States, 2009-2013, by Consumers' Demographic Characteristics. *Research Data Reports Paper* 15-6 (2015).

[15] Robert Cull, Vivien Foster, Dean Jolliffe, Daniel Lederman, Veerappan Malarvizhi, Davide S Mare, and Davide Salvatore Mare. 2023. *Digital Payments and the COVID-19 Shock.* World Bank.

[16] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347, 6221 (2015), 536–539.

[17] Ben Derrick, Deirdre Toher, and Paul White. 2016. Why Welch's test is Type I error robust. *The Quantitative Methods in Psychology* 12, 1 (2016).

[18] Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C González. 2018. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nature communications* 9, 1 (2018), 1–8.

[19] Mary-Alice Doyle, Chay Fisher, Ed Tellez, Anirudh Yadav, et al. 2017. RDP 2017-04: How Australians Pay: Evidence from the 2016 Consumer Payments Survey. *Reserve Bank of Australia Research Discussion Papers* July (2017).

[20] Angelo Duarte, Jon Frost, Leonardo Gambacorta, Priscilla Koo Wilkens, and Hyun Song Shin. 2022. Central banks, the monetary system and public payment infrastructures: lessons from Brazil's Pix. *Available at SSRN 4064528* (2022).

[21] Federal Statistical Office. 2022. *General Classification of Economic Activities (NOGA).* https://www.bfs.admin.ch/bfs/en/home/statistics/industry-services/nomenclatures/noga.html

[22] Federal Statistical Office. 2022. *Highest completed educational level.* https://www.bfs.admin.ch/bfs/en/home/statistics/population/migration-integration/integration-indicators/all-indicators/education-training/level-education-completed.html

[23] Federal Statistical Office. 2022. *Räumliche Typologien.* https://www.bfs.admin.ch/bfs/de/home/statistiken/querschnittsthemen/raeumliche-analysen/raeumliche-gliederungen/raeumliche-typologien.html

[24] Federal Statistical Office. 2022. SAKE in Kürze 2021 - Schweizerische Arbeitskräfteerhebung.

[25] Federal Statistical Office. 2022. *Ständige Wohnbevölkerung nach Postleitzahl, Staatsangehörigkeitskategorie, Geschlecht, Altersklasse und Zivilstand, 2010-2021.* https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases.assetdetail.su-d-01.02.03.07.html

[26] Laura Felber and Simon Beyeler. 2023. Nowcasting economic activity using transaction payments data. (2023).

[27] Marie-Hélène Felt, Fumiko Hayashi, Joanna Stavins, and Angelika Welte. 2020. Distributional effects of payment card pricing and merchant cost pass-through in the United States and Canada. *Federal Reserve Bank of Kansas City Working Paper* 20-18 (2020).

[28] John W Galbraith and Greg Tkacz. 2015. *Nowcasting GDP with electronic payments data.* Technical Report. ECB Statistics Paper.

[29] Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. 2022. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234* (2022).

[30] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. 2006. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems* 19 (2006).

[31] George Kapetanios, Fotis Papailias, et al. 2021. *UK Economic Conditions during the Pandemic: Assessing the Economy using ONS Faster Indicators*. Technical Report. Economic Statistics Centre of Excellence (ESCoE).

[32] Tae Kyun Kim. 2015. T test as a parametric statistic. *Korean journal of anesthesiology* 68, 6 (2015), 540–546.

[33] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* 110, 15 (2013), 5802–5805.

[34] Sergei Koulayev, Marc Rysman, Scott Schuh, and Joanna Stavins. 2016. Explaining adoption and use of payment instruments by US consumers. *The RAND Journal of Economics* 47, 2 (2016), 293–325.

[35] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[36] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2018), 2346–2363.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[39] Laura Schelenz, Ivano Bison, Matteo Busso, Amalia De Götzen, Daniel Gatica-Perez, Fausto Giunchiglia, Lakmal Meegahapola, and Salvador Ruiz-Correa. 2021. The theory, practice, and ethical challenges of designing a diversity-aware platform for social relations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 905–915.

[40] Stanislav Sobolevsky, Emanuele Massaro, Iva Bojic, Juan Murillo Arias, and Carlo Ratti. 2017. Predicting regional economic indices using big data of individual bank card transactions. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1313–1318.

[41] Masashi Sugiyama. 2015. *Introduction to statistical machine learning*. Morgan Kaufmann.

[42] Swiss National Bank. 2018. Survey on payment methods 2017: Survey on payment behaviour and the use of cash in Switzerland. (2018).

[43] Swiss National Bank. 2021. Survey on payment methods 2020: Survey on payment behaviour and the use of cash in Switzerland. (2021).

[44] Swiss National Bank. 2023. Survey on payment methods 2022: Survey on payment behaviour and the use of cash in Switzerland. (2023).

[45] Victoria Vickerstaff, Rumana Z Omar, and Gareth Ambler. 2019. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC medical research methodology* 19, 1 (2019), 1–13.

[46] Alice Zheng and Amanda Casari. 2018. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

A Appendices

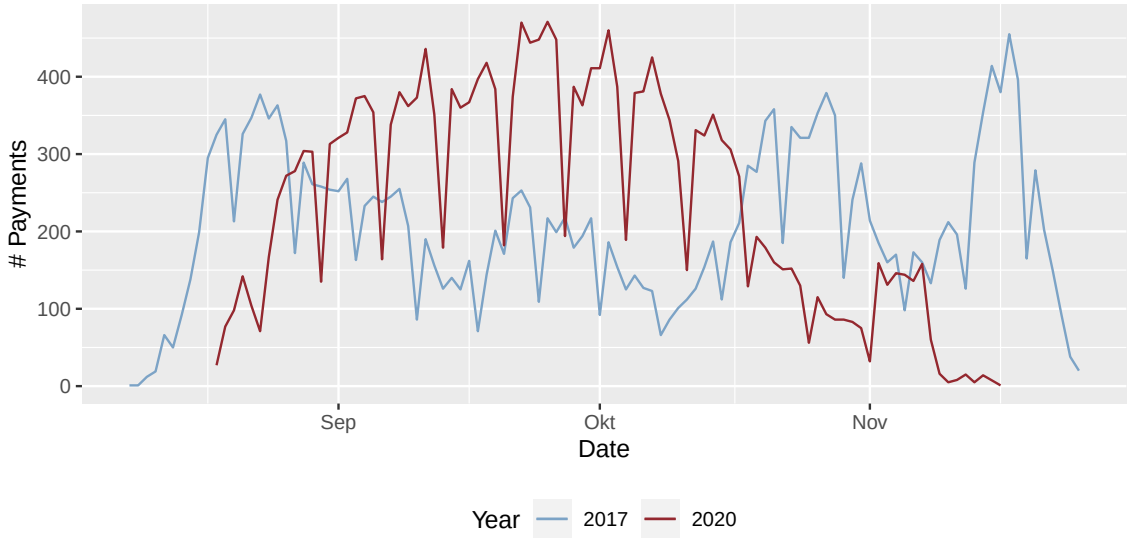


Fig. 9. Number of payments per day for the 2017 and 2020 payment surveys.

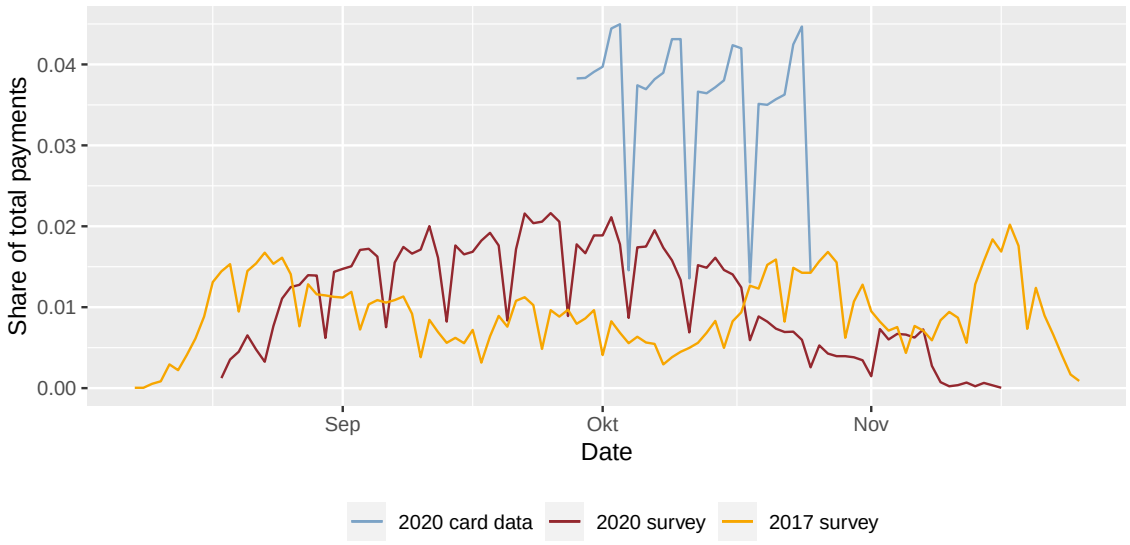


Fig. 10. Normalized distribution of the number of payments in the card data and in the payment surveys.

Table 5. Mapping between NOGA codes, payment attributes and survey-defined payment purposes.

Survey Location	Card Data NOGA Codes	Payment Attribute	Resulting Category
Convenience store	47		Retail
Durable products	47		Retail
Petrol stations	47		Retail
Postal shopping	47		Retail
Travelling	55		Travelling
Restaurant/bar	56		Restaurant/bar
Leisure	93		Leisure
Other services	96/49		Services
Payments to individuals	96/49		Services
Household services	96/49		Services
E-Commerce	Other	Online	E-Commerce
Vending machines	Other	Unattended POS	Vending machines
Other	Other		Other
Governmental services	Other		Other
Donations	Other		Other

Table 6. Summary table of the feature extraction.

Feature set **# features**	**Temporal features** **# features**	**Payment instrument features** **# features**	**Payment location features** **# features**
Extensive	Numeric week of the year. Number of payments, mean and sd[1] payment value for: Monday/ Tuesday/ Wednesday/ Thursday/ Friday/ Saturday/ Sunday	Number of payments, mean and sd[1] payment value for: Cash / Credit Card / Credit Card, CL / Debit Card / Debit Card, CL / Prepaid Card / Prepaid Card, CL / Mobile/Online Payment / Merchant Specific / Online Banking / Teller/ Others	Number of payments, mean and sd[1] payment value for: Restaurant, bars, takeaway / Convenience store / Durable products / Petrol stations / Vending machines / E-Commerce / Postal shopping Payments to individuals / Household services / Governmental services / Other services / Leisure / Donations / Travelling / Other locations
103 features	22 features	36 features	45 features
Simplified	Numeric week of the year. Number of payments, mean and sd[1] payment value for: Week	Number of payments, mean and sd[1] payment value for: Cash / Credit Card / Debit Card / Prepaid Card / Mobile/Online Payment / Merchant Specific / Online Banking / Teller/ Contactless / Others	Number of payments, mean and sd[1] payment value for: Restaurant, bars, takeaway / Convenience store / Durable products / Petrol stations / Vending machines / E-Commerce / Postal shopping Payments to individuals / Household services / Governmental services / Other services / Leisure / Donations / Travelling / Other locations
79 features	4 features	30 features	45 features
Card data emulated	Numeric week of the year. Number of payments, mean and sd[1] payment value for: Week	Card type dummy (debit, credit, prepaid or mobile). Number of payments, mean and sd[1] payment value for contactless	Number of payments, mean and sd[1] payment value for: Convenience store / Restaurant, bars, takeaway / Vending machines / E-Commerce / Travelling / Leisure / Services / Other locations
34 features	4 features	6 features	24 features

sd = standard deviation, CL = Contactless

Table 7. Summary statistics per binary sociodemographic group based on the 2020 survey.

Demographics	$\#_{payments}$	$mean_{value}$	sd_{value}	Demographics	$\#_{payments}$	$mean_{value}$	sd_{value}
Below 25	7.80	28.62	36.12	25 above	10.93	55.80	79.31
Above 65	10.83	61.35	88.06	Below 65	10.39	48.86	68.31
In education	7.43	30.23	37.66	Not in education	10.89	54.89	77.98
Employed	11.05	52.57	76.28	Not employed	9.76	51.54	69.62
Female	10.32	52.25	69.22	Male	10.69	52.02	77.77
Married	10.74	59.12	85.01	Not married	10.25	44.57	61.01
Low fin index	8.91	41.09	46.19	Not low fin index	10.76	53.95	78.08
High fin index	11.42	66.28	104.37	Not high fin index	10.36	49.81	68.54
Primary educated	7.93	35.09	38.21	Not low educated	10.77	53.70	76.71
Tertiary educated	11.31	55.94	80.71	Not high educated	9.90	49.03	67.52

Table 8. Top 5 most indicative features as measured by Student's t-tests for each sociodemographic group in the 2017 survey.

	feature	t_stat	Diff	[95% CI Diff]	CD		feature	t_stat	Diff	[95% CI]	CD
Below 25	$\overline{value}$ Petrol stations	8.47****	-10.93	[-13.47, -8.4]	-0.41	Above 65	#Debit Card	7.23****	-1.12	[-1.42, -0.81]	-0.39
	$\overline{value}$ Retail	7.95****	-22.81	[-28.44, -17.19]	-0.26		#E-Commerce	6.87****	-0.12	[-0.16, -0.09]	-0.25
	$\overline{value}$ Cash	7.8****	-13.28	[-16.61, -9.94]	-0.23		$\overline{value}$ E-Commerce	4.67***	-12.44	[-17.66, -7.22]	-0.14
	$\sigma(value)$ Retail	7.03****	-23.14	[-29.6, -16.69]	-0.19		#Vending machines	4.48***	-0.29	[-0.42, -0.16]	-0.22
	$\overline{value}$	6.65****	-22.23	[-28.8, -15.67]	-0.28		#Donations	4.41**	0.12	[0.07, 0.18]	0.38
Student	$\overline{value}$ Petrol stations	11.63****	-14.08	[-16.46, -11.69]	-0.53	Employed	$\overline{value}$ Petrol stations	8.87****	10.27	[8, 12.54]	0.39
	#Donations	10.41****	-0.08	[-0.1, -0.07]	-0.26		#Debit Card	6.82****	0.88	[0.63, 1.13]	0.31
	$\overline{value}$ Debit Card	10.27****	-26.62	[-31.72, -21.53]	-0.38		#Petrol stations	6.15****	0.31	[0.21, 0.41]	0.27
	$\overline{value}$ Cash	8.14****	-14.84	[-18.41, -11.26]	-0.26		#Payments	5.67****	1.49	[0.98, 2.01]	0.26
	$\overline{value}$	8.01****	-25.91	[-32.27, -19.55]	-0.32		#Restaurant/bar	5.58****	0.80	[0.52, 1.08]	0.24
Female	#Restaurant/bar	8.96****	-1.31	[-1.6, -1.02]	-0.41	Married	$\overline{value}$ Petrol stations	4.9****	6.14	[3.68, 8.61]	0.23
	#Petrol stations	6.87****	-0.35	[-0.45, -0.25]	-0.31		$\overline{value}$	4.1**	15.64	[8.16, 23.13]	0.20
	$\overline{value}$ Petrol stations	5.58****	-6.76	[-9.13, -4.38]	-0.25		$\overline{value}$ Debit Card	3.98**	13.16	[6.66, 19.65]	0.19
	#Other services	4.98****	0.19	[0.12, 0.27]	0.22		#Credit card	3.34	0.22	[0.09, 0.35]	0.15
	$\sigma(value)$ Petrol stations	4.35**	-1.87	[-2.71, -1.02]	-0.20		$\sigma(value)$ Restaurant/bar	3.16	4.05	[1.54, 6.57]	0.16
Low fin	$\overline{value}$ Petrol stations	8.87****	-10.64	[-13, -8.29]	-0.41	High fin	#Travelling	5.83****	-0.03	[-0.03, -0.02]	-0.15
	#Credit card	6.48****	-0.42	[-0.54, -0.29]	-0.29		$\overline{value}$ Travelling	4.08**	-4.90	[-7.26, -2.54]	-0.10
	#Restaurant/bar	6.09****	-0.94	[-1.25, -0.64]	-0.29		#Credit card	3.45	0.91	[0.39, 1.44]	0.64
	$\overline{value}$ Restaurant/bar	5.67****	-7.88	[-10.61, -5.16]	-0.26		$\overline{value}$ Petrol stations	2.82	13.48	[3.95, 23.01]	0.51
	$\sigma(value)$ Restaurant/bar	5.46****	-6.20	[-8.42, -3.97]	-0.24		$\overline{value}$ Debit Card	2.28	19.34	[2.42, 36.27]	0.26
Primary edu.	$\overline{value}$ Petrol stations	6.52****	-8.88	[-11.56, -6.21]	-0.33	Tertiary edu.	#Credit card	5.45****	0.56	[0.36, 0.77]	0.41
	#Credit card	5.84****	-0.35	[-0.46, -0.23]	-0.25		#Payments	5.04****	1.87	[1.14, 2.61]	0.32
	$\overline{value}$ Debit Card	5.78****	-19.10	[-25.6, -12.6]	-0.27		#Debit Card	4.04**	0.67	[0.35, 1]	0.23
	$\overline{value}$ Restaurant/bar	5.21****	-7.69	[-10.59, -4.79]	-0.27		#Restaurant/bar	3.88**	0.70	[0.34, 1.05]	0.21
	$\overline{value}$ Credit card	5.05****	-16.24	[-22.56, -9.93]	-0.16		#contactless	3.52	0.36	[0.16, 0.56]	0.26

Table 9. Results including accuracy, precision and recall metrics per sociodemographic group for Random Forest, XG Boost, Ada Boost.

Demographic	$\#_{samples}$	**Random Forest**			**XG Boost**			**Ada Boost**		
		$\overline{A}$	$\overline{P}$	$\overline{R}$	$\overline{A}$	$\overline{P}$	$\overline{R}$	$\overline{A}$	$\overline{P}$	$\overline{R}$
In education	(468)	83.11	81.36	86.76	82.90	81.13	86.76	83.51	82.21	87.19
Below 25	(566)	82.00	83.41	80.27	81.09	81.26	81.26	80.74	80.97	80.22
Above 65	(1100)	71.00	69.90	74.18	71.64	71.68	71.82	69.91	69.63	70.73
In retirement	(1136)	70.96	71.49	70.26	70.60	71.25	69.20	71.39	71.86	70.79
Primary edu.	(376)	70.42	72.88	66.52	67.73	69.24	65.35	67.79	67.97	68.65
High fin.	(590)	67.30	66.43	71.22	65.25	67.21	60.02	64.91	67.47	58.31
Married	(2016)	65.92	64.47	71.43	65.28	65.82	63.80	65.13	65.79	63.31
Employed	(1764)	64.92	65.66	63.15	64.52	64.19	65.99	64.12	63.31	67.46
Female	(2094)	63.70	63.77	63.51	63.89	64.44	62.18	64.90	65.04	64.66
Low fin.	(558)	64.37	63.63	67.79	61.81	60.87	66.30	60.24	58.50	70.30
Tertiary edu.	(1796)	61.92	62.93	58.69	60.30	61.44	56.12	60.80	62.37	54.91
Countryside	(704)	58.67	58.18	60.21	57.52	56.83	62.73	56.11	55.56	60.53
City	(1628)	55.28	55.79	50.61	54.97	55.10	52.33	52.22	52.32	50.37
Swiss citizenship	(598)	55.53	55.21	60.53	56.21	56.20	57.87	54.54	55.09	56.14

(a) Trains on 2017 and tests on 2020

(b) Trains on 2020 and tests on 2017

Fig. 11. Sociodemographic inference results when training on one year and testing on the other year. The results when training and testing on the same year are provided as a baseline performance measure.