CRC Credit Research Centre

CRC Working Papers

The Double-edged Sword Of Big Data And Information Technology For The Disadvantaged: A Cautionary Tale From Open Banking

By Savina D. Kim, Galina Andreeva, Michael Rovatsos

July 2023

Keywords: Fairness, Discrimination, Financial vulnerability, Credit scoring, FinTech, Open Banking, Alternative credit data

THE UNIVERSITY of EDINBURGH

Contact
Credit Research Centre
University of Edinburgh Business School
29 Buccleuch Place
Edinburgh, EH8 9JS
credit.research.centre@ed.ac.uk

THE DOUBLE-EDGED SWORD OF BIG DATA AND INFORMATION TECHNOLOGY FOR THE DISADVANTAGED: A CAUTIONARY TALE FROM OPEN BANKING

A PREPRINT

Savina D. Kim *
University of Edinburgh
Edinburgh, United Kingdom

Galina Andreeva
University of Edinburgh
Edinburgh, United Kingdom

Michael Rovatsos
University of Edinburgh
Edinburgh, United Kingdom

July 20, 2023

ABSTRACT

This research article analyses and demonstrates the hidden implications for fairness of seemingly neutral data coupled with powerful technology, such as machine learning (ML), using Open Banking as an example. Open Banking has ignited a revolution in financial services, opening new opportunities for customer acquisition, management, retention, and risk assessment. However, the granularity of transaction data holds potential for harm where unnoticed proxies for sensitive and prohibited characteristics may lead to indirect discrimination. Against this backdrop, we investigate the dimensions of financial vulnerability (FV), a global concern resulting from COVID-19 and rising inflation. Specifically, we look to understand the behavioral elements leading up to FV and its impact on at-risk, disadvantaged groups through the lens of fair interpretation. Using a unique dataset from a UK FinTech lender, we demonstrate the power of fine-grained transaction data while simultaneously cautioning its safe usage. Three ML classifiers are compared in predicting the likelihood of FV, and groups exhibiting different magnitudes and forms of FV are identified via clustering to highlight the effects of feature combination. Our results indicate that engineered features of financial behavior can be predictive of omitted personal information, particularly sensitive or protected characteristics, shedding light on the hidden dangers of Open Banking data. We discuss the implications and conclude fairness via unawareness is ineffective in this new technological environment.

Keywords Fairness · Discrimination · Financial vulnerability · Credit scoring · FinTech · Open Banking · Alternative credit data

1 Introduction

Advances in information technology coupled with the availability of new data can fundamentally change the world for the better. Indeed, there are multiple examples of this already happening, from wearable devices and smart digital infrastructure helping improve outcomes of chronic disease patients [Ghose et al., 2022] to online crowdfunding platforms providing small businesses with an alternative financing channel [Luo et al., 2022]. Nevertheless, it is important to be aware of the unintended side-effects of any innovation, especially in relation to disadvantaged segments of the population. In this paper, we illustrate the potential risks of one such technology in the sector of household finance; specifically, we highlight the hidden associations of seemingly neutral Open Banking (OB) transactions with sensitive and protected personal characteristics of the financially vulnerable. To the best of our knowledge, this is the first study where solid evidence for this is given using real OB data.

With growing enthusiasm around machine learning (ML) and alternative credit data over the last few years, the implementation of the second Payment Services Directive (PSD2 or Open Banking in the UK[2]) is a game changer,

*S. Kim is affiliated with the University of Edinburgh Business School and the Centre for Technomoral Futures at the Edinburgh Futures Institute, United Kingdom

[2]On October 27, 2022, the Consumer Financial Protection Bureau (CFPB) announced the introduction of Open Banking in the USA. This announcement extends a worldwide trend in opening financial transactions data which makes our investigation even more timely and relevant.

particularly in credit risk assessment [Remolina, 2019]. It enables easier accessibility to dynamic, real-time consumer data in practice, with bank transaction data being of particular interest. However, this rapid data shift and fast-paced development of algorithmic decision-making has simultaneously raised several social and ethical concerns. This is worsened by the fact that regulations "have not kept pace with modern Big Data capabilities" [Wolkowitz and Parker, 2015, p. 24] where the use of expanded data types is still porously regulated and largely unbounded. Questions arise regarding the magnitude of power of OB data in customer profiling, and whether one should be cautious from the standpoint of fairness and equality principles [Hiller and Jones, 2022].

Fairness, and especially bias, in data and algorithmic decision-making has recently received a lot of attention from information science academics and practitioners [Bellamy et al., 2019, Chouldechova and Roth, 2018, Lessmann et al., 2015, Mehrabi et al., 2019, Corbett-Davies et al., 2017]. This paper extends this line of inquiry by examining the associations of sensitive and legally protected personal characteristics with seemingly neutral financial transactions. In this study, we look to better understand the risks concealed in novel types of data which may go unnoticed. In doing so, we simultaneously provide a warning to all data modelers, and especially to lenders who are too hastily taking an "all data is credit data" [Aitken, 2017] approach without proper caution. We investigate these risks in the context of financial vulnerability (FV), a common denominator across those in need of credit as well as being an urgent, global concern. Due to the unprecedented economic impact brought by the COVID-19 shock and further exacerbated by double-digit inflation and a looming energy crisis, a sharp rise to living costs has dramatically accelerated financial distress worldwide. In the UK alone, the number of adults with low financial resilience increased by 3.5 million between March and October of 2020, according to the Financial Lives survey [Financial Conduct Authority, 2021] with similar conclusions made by Lowell's Vulnerability Index which indicated a 11% rise in financial vulnerability (FV) across the pandemic [Braga et al., 2021]. A survey in the US revealed significant inequality in financial impact of COVID-19, where those most financially vulnerable prior to the pandemic faced even greater financial strain [Bruce et al., 2022].

Surprisingly, there is no research connecting FV with the objective financial data, as previous limited research relied on self-reported FV measures from surveys. To fill this gap, we provide a comprehensive investigation of FV dimensions and their drivers using information derived from bank transaction data. In doing so, we demonstrate the power of OB data in profiling FV, which can be used for social good to help the disadvantaged segments. On the other hand, the same power can be abused e.g., by predatory lenders targeting customers who cannot afford credit repayments [Jones et al., 2020].

The analysis is performed on a large OB dataset containing 180 million bank transaction records from up to 100,000 individuals in the UK sourced from a social FinTech lender. This unique dataset also contains some socio-demographic variables, including gender, which is not normally collected by lenders, since it is one of the characteristics not allowed for making credit-granting decisions.[3] From financial transactions we can infer another protected characteristic – disability, and two additional ones that can be considered sensitive, namely whether the account-holder is a carer and if s/he has children.

Drawing from regulatory guidance [Financial Conduct Authority, 2015], we propose six binary data-driven FV indicators (FVI) that can be used to identify vulnerable customers. We then generate a wide range of financial behaviors from financial management to financial inclusion that may be associated with FV as suggested by regulations and previous research. Knowledge of which behaviors may lead to FV can help in designing appropriate interventions. Bivariate Pearson linear correlation reveals statistically significant associations not only between FVI and financial behaviors, but also with protected characteristics. We use state-of-the-art machine learning algorithms to predict FVIs from financial behaviors and compare predictive accuracy as measures of successful performance. This approach has an advantage over Pearson's correlation coefficient in capturing complex non-linear associations with multiple predictors. We also predict protected characteristics from financial behaviors and observe very high predictive accuracy, signaling powerful associations, and demonstrating the power of OB data.

We then illustrate the potential biases that may arise through unawareness of these correlations. We show how segmentation and clustering (a popular tool in marketing and customer management) can implicitly capture sensitive and protected characteristics, even when they are not included in the data and its analysis. Furthermore, we reveal that clusters are characterized by the combination of several sensitive characteristics. Therefore, caution should be taken in making decisions not only with regards to a particular individual sensitive or protected characteristic, but also in terms of their combinations.

The contributions of this paper can be summarized on two levels. First, on a more general level, we contribute to the body of research on responsible and ethical AI through the lens of a real-world scenario and a practical application with lessons learned for fair modeling practices. The hidden connections with protected and sensitive characteristics can be present in any data across various applications, thus modelers and end users should be aware of this risk. We also extend

[3]The data provider collects gender for Know Your Customer (KYC) purposes, and does not use it for credit decisions.

the discourse on data biases by looking at the new types of alternative credit data that have not been analyzed previously, particularly granular transaction data. Second, at the context-specific level, we make several important contributions by quantifying the dimensions and drivers of FV. This is accomplished by demonstrating the predictive power of OB data in modeling FV and by proposing a data-driven segmentation of financial behaviors for enhanced applicant profiling.

The remainder of this paper is structured as follows. Section 2 summarizes the relevant studies, including the impact of Open Banking, an examination of fairness concerns in credit systems and FV assessment methods. Section 3 introduces the FV indicators, followed by the empirical results of the ML-based prediction and clustering methods in Section 4. The final Section 5 concludes the paper with a discussion of the findings, its implications for public policy makers and practitioners, and limitations which offer promising opportunities for future research.

2 Related Work

2.1 Open Banking

Open Banking refers to a regime in which banks provide access to customer financial transactions in secure, digitalized form – with customers' consent – to authorized third-party service providers such as FinTech companies. Standardized open application programming interfaces (APIs) are used by third parties to deliver services to customers using their own data, ranging from money management applications and financial product comparisons to applications for loans or mortgages. This has revolutionized financial services, pushing the boundaries of traditional credit risk assessment (i.e., credit scores), which often fail to identify more nuanced behaviors leading up to financial difficulty and rely on static, outdated data pulls rendering them opaque and error prone [O'Leary et al., 2021]. This, coupled with the growing power of computational tools and reduction in data storage costs, has opened a new frontier for extracting inference from big data [Côrte-Real et al., 2017].

In particular, transaction records can reflect a consumer's risk profile on the basis of historical and current financial habits and preferences. Vissing-Jorgensen [2012] argues there is informative value in what consumers buy when predicting default risk; for example, spending on entertainment such as magazines or toys. Customers' personal information and aggregated transaction history on annual and monthly bases have also been shown to be effective [Zhang et al., 2018] along with the balance of checking and savings accounts and cash inflows and outflows [Khandani et al., 2010]. Even grocery shopping data has been shown to be informative when predicting credit card repayment behaviors [Lee et al., 2021]. However, these studies focus on only creditworthiness, skipping over its major determinants such as financial health, stability and, correspondingly, vulnerability. Our study seeks to close this gap. To summarize, alternative data enables financial institutions to gain a more holistic view of an individual's financial health and can help overcome the limitations of traditional methods; however, its intrusiveness simultaneously increases potential for individual harm. Therefore, how to guarantee consumer priorities such as fairness is still an open question, noting the minimal research on Open Banking data in relation to fairness considerations needed for its ethical use. These concerns are discussed in detail next.

2.2 Discrimination in Credit and Algorithmic Bias

One major concern around the use of alternative data for predicting individual levels of risk is enabling predatory lenders to identify vulnerable groups more easily and further perpetuating the harmful cycle of discrimination [Hiller and Jones, 2022]. In other words, the same technology designed to make the "credit invisibles" of contemporary financial markets visible, is the same technology with the ability to precisely identify them [Jones et al., 2020]. Prominent examples of this occurring are the targeting of minority borrowers for high-interest, subprime loans by fringe lenders making it harder for them to build a strong history of repayment; or the historic practice of redlining, resulting in a vicious cycle of restricted access which dramatically impacts communities of color [Hiller and Jones, 2022]. According to sociological investigations by Fourcade and Healy [2017], segmentation methods increase social stratification and subjectification in terms of access to consumer credit in the American market, negatively impacting the life chances of select individuals. Similar conclusions were drawn by others when analyzing the technological disruption of the fringe finance infrastructure [Langevin, 2019]. This risk is magnified by findings which show that offering financial incentives (e.g., lower interest rates) drives many to disclose personal information with minimal consideration of the possible consequences to such data sharing [Acquisti et al., 2013, Norberg et al., 2007], risking the transition of unsound fringe lenders into mainstream lenders. Packin [2016] also explore potential consequences of emerging social credit systems, which authorize the use of highly personal information in return for better interest rates. They sound the alarm claiming direct and derivative harms to loan seekers regarding privacy, social segregation and due process violations derived from unsupervised ML.

Facially neutral features may also be highly correlated with a protected characteristic (e.g., race or disability). A predictive model works by capturing all the features characterizing an event which is then utilized to make predictions; however, certain features may not only characterize the intended event but also inform another phenomena or class, also known as proxies [Veale and Binns, 2017]. For example, factors such as educational background, wealth, work history, geography or even where one goes grocery shopping can be used to infer an individual's race and thus exploited without proper regulation [Hurley and Adebayo, 2017]. Furthermore, because fitting the majority population is more important for reducing overall modelling error, this leads to different (and often higher) distribution of errors for the minority population, who are thus systemically handicapped to begin with [Chen et al., 2018, Chouldechova and Roth, 2018]. ML techniques are designed to fit the data therefore it is expected that they will replicate and amplify any bias already existing in the data; we have no reason to expect them to remove it. Therefore, utilizing unbalanced data can be particularly harmful towards underrepresented (and underfinanced) groups, which is often the case in financial services [Rovatsos et al., 2019].

Alongside embedded biases in credit scoring models, economic hardships are also reported at higher rates for racial and ethnic minorities, making them particularly a target. Black (73%) and Hispanic adults (70%) reported that they lacked emergency funds to cover three months of expenses (compared to 47% White adults), and they more often reported (48% and 44%, respectively) that they would be unable to fully pay their bills during the recent COVID-19 pandemic, a real-world example of financial shock [Lopez et al., 2020].

2.3 Financial Vulnerability

FV, often used interchangeably with financial fragility, distress, debt burden, and overindebtedness, refers to an individual's ability to manage daily finances, their resilience to economic shocks (e.g., unexpected rent increase or medical expense), and their capacity to pursue financial opportunities. Conceptually, it is a multi-faceted concept making it difficult to define and thus quantify [Perrig-Chiello et al., 2016]. To help financial institutions and policymakers identify and support these 'at-risk' individuals, recent studies have looked to define formative measures of FV in hopes of attenuating its negative impact on financial outcomes.

On an empirical basis, analysis of objective financial outcomes, or those straightforwardly derived from bank account records, are common. These include assessing one's ability to pay their debts, such as debt-to-income ratio [Costa and Farinha, 2012] or timely repayment, when debt is in arrears for more than 90 days [Il et al., 2016]. Other indicators measure income, expenses, and wealth levels [Ampudia et al., 2016], including whether cash flow can cover basic living costs such as utility bills [Bridges and Disney, 2004]. Being unable to take a vacation, going out for a meal with friends, or enjoying leisure activities are also considered [Worthington, 2006]. Subjective approaches using self-reported surveys are also used to capture individuals' own perception of their financial situation. Whereas more difficult to scale and interpret, they capture a unique aspect of consumer welfare, such as one's anxiety over personal finances, which purely objective financial data cannot. For example, Lusardi et al. [2011] ask individuals how confident they are that they could come up with $2000 in 30 days to face an unexpected need. Policy-making bodies and regulators have also begun to pay greater attention to the subjective dimension, evidenced by the Financial Lives surveys [Financial Conduct Authority, 2017, 2021] and the measures for financial wellbeing [Consumer Financial Protection Bureau, 2015]. Other research streams have examined the impact of behavioral patterns on FV likelihood, including money management skills, personal savings orientation, consideration of future consequences [Rustichini et al., 2016] as well as psychological characteristics [Gladstone et al., 2019]. It is argued that financial illiteracy and lack of self-control are major determinants [Gathergood, 2012, Lusardi, 2019], evidenced by high levels of overindebtedness in the most vulnerable individuals with multiple sources of (often unsecured) debt. It seems that behavioral biases, such as impulsivity, may lead to spending beyond one's means as they are unable to resist instant gratification [DeHart et al., 2016]. On the other hand, to meet necessary living costs under income shocks, individuals may also have no choice but to rely on resources outside of their savings, pensions or benefits, namely, consumer credit. However, vulnerable individuals often lack or have insufficient credit history resulting in financial exclusion and deprivation, particularly from traditional financial institutions [Brevoort et al., 2016]. We rely on these findings and insights when defining features, indicators, and determinants of FV as described in the next section.

3 Data and Methods

3.1 Data Description

We leverage a new, proprietary dataset provided by a UK-based social lender that primarily lends to individuals working in the public sector with loan sizes typically ranging between £500 to £1,000. The lender is unique in that they assess applicants via Open Banking during the affordability check procedure, excluding the need for a credit score. Using a third-party Open Banking API, the lender aggregates transactions across the applicant's current account. The data was collected in February 2022 and focuses on approximately 100,000 applicants who have applied for a loan in the previous two years, yielding a dataset of over 180 million transactions. All applicant information is anonymized and includes demographic variables as illustrated in Figure 1.

Historical bank transaction records of each applicant include transaction amount in British pounds (GBP), date, description, classification, category and remaining account balance. The data provider uses a third-party algorithm to categorize transactions into 56 categories (e.g., groceries & housekeeping, earnings, etc.) with transaction references providing additional color on the merchant or service provided. To ensure a sufficient level of data per applicant and capture those actively using their account, we retain applicants who exhibit complete transaction details for a minimum observation period of six months and with at least ten transactions per month.

Variable	Summary	Data Type	Example Values
Age	Age ranging from 18-60+, with largest proportion (19%) of applicants being 31 - 35 years old, followed by 26-30 and 36-40 year olds (each 17%)	Categorical (10 buckets)	18 - 21, 22 - 25, 26 - 30...60+
Sex	74% female and 26% male	Indicator	F, M
Employment length	Self-reported length of employment with largest proportion being over 4 years (44%); 2 - 4 years (24%); 1 - 2 years (14%); 6 - 12 months (12%); Less than 6 months (6%)	Categorical (5 buckets)	Less than 6 months, 6 - 12 months…Over 4 years
Residential status	Self-reported residential status with largest proportion being Private Tenant (32%); Housing Association Tenant (21%); Living with Parents (17%); Owner Occupier (15%); Council Tenant (14%)	Categorical (5 buckets)	Council tenant, owner occupier, living with parents

Figure 1: Details of variables used from applicants' demographic information data.

3.2 Financial Vulnerability Indicators

There is no official or universally accepted definition of FV. Since the data is from the UK context, we refer to the Financial Conduct Authority's (FCA) guidance on financial difficulties [Financial Conduct Authority]. We consider key indicators of being 'financially in difficulty' or having 'low financial resilience' which include having insufficient funds in their account, being over-indebted, having low or erratic incomes or low savings and being unable to withstand an unexpected increase in monthly expenses such as rent. In line with the FCA, six target binary variables/key FV indicators are derived to measure applicants' status and include the following.

- *(1) Financial shock withstanding* (48.3% of applicants): When an applicant is unable to withstand the impact of an unexpected expenditure of £100 on their account for more than 50% of the months throughout their account history as included in the training or test samples. This is computed by referencing the average monthly median of an applicant's account balance, where a median value less than £100 would suggest the applicant would struggle to sustain a shock. Since we only have access to transaction data from current/checking accounts, we are unable to view savings amounts, but this is likely to be minimal given only 4% of applicants show evidence of payments into the category of *savings and investments*. Additional threshold levels, including when individuals can withstand financial shock 0% of all months and thus extremely financially vulnerable (*"financial shock withstanding never"* comprising 18.4% of applicants) and when individuals can withstand financial shock 100% of all months and thus financially healthy (*"financial shock withstanding always"* comprising 19.0% of applicants), are assessed to represent more extreme cases. This enables us to compare across different intensities of FV and determine if select behaviors are monotonic.
- *(2) Insolvent* (4.0% of applicants): When an applicant has made at least one or more payments under the transaction category of debt management and insolvency in their history. This indicates an applicant has not been able to pay their debts when they were due and therefore overindebted.
- *(3) Insufficient disposable income* (11.6% of applicants): When an applicant has less than or equal to £100 in average monthly disposable income. To determine whether individuals can cope with a financial shock, we assess whether they can afford their necessary expenditures or would be deprived.

- *(4) Overdraft* (67.4% of applicants): When an applicant has at least one or more days in overdraft (OD) per month for more than 50% of the months throughout their account history. OD indicates insufficient funds in the account. Other thresholds where individuals are in OD 0% of months and thus financially healthy (*"overdraft never"* comprising 6.9% of applicants) and individuals are in OD 100% of months and thus extremely financially vulnerable (*"overdraft always"* comprising 35.1% of applicants) are also assessed.
- *(5) Returned direct debits* (28.3% of applicants): When an applicant has at least one or more returned direct debits (RDD) on average per month. This value indicates insufficient funds resulting in a rejection of a pre-arranged payment by the bank.
- Another feature of interest, which does not directly imply FV, however, may be related to managing financial matters, includes being a *(6) Gambler* (20.2% of applicants). An applicant is considered a gambler if they have spent £100 or more on average per month on gambling expenditures. We propose that it is the compilation of indicators which constitute the umbrella term FV.

3.3 Protected and Sensitive Attributes

Central to this study are legally protected characteristics given their importance in anti-discrimination laws, also referred to as fair lending laws, such as the Fair Credit Reporting Act (FCRA), Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA) in the United States, the Equality Act in the United Kingdom and similar consumer protection directives in the European Union with Articles 12 and 13 of the EC Treaty which prohibit and provide measures to combat discrimination, respectfully. These make up the most significant legal instruments governing consumer credit scoring regulation which outline processes by which fairness should be met, most notably by prohibiting discrimination by racial or ethnic origin, nationality, religion or belief, sex, age, disability, or sexual orientation, and being used as basis for decision-making.

We also consider some additional characteristics that may be seen as sensitive, especially if used in decision-making. These characteristics are derived from the FCA guidance, which describes a list of risk factors common to vulnerable consumers who, due to personal circumstances, are more susceptible to financial detriment [Financial Conduct Authority, 2015]. These risk factors are also highlighted in consumer and public policy research [Anderson et al., 2018, Griffiths and Harmon, 2011, Mos, 2011] and include: low education and financial literacy, physical disabilities, severe or long-term illnesses, mental health issues, low income, high debt, caring responsibilities, being either "young" or "old," lack of English language skills, and impactful changes in personal circumstances, such as a divorce, death of a spouse, or a redundancy. Not every individual falling into one or more of these categories will necessarily experience FV, however, these factors are expected to increase the susceptibility of entering a financially vulnerable state as well as experiencing the severity of its consequences.

Taking both risk factors and protected characteristics into consideration, our work examines select attributes that are available in our data. We refer to these as 'socio-demographic profile' features which describe an applicant and are commonly considered influential factors in their behavior and financial status. The lending company obtains socio-demographic characteristics at origination, including gender and age, alongside what we infer as sensitive attributes. These include whether an applicant has a disability (8% of applicants), which is determined using disability benefit payments as a proxy, and attributes such as having a child (37.5% of applicants) or being a carer (2.5% of applicants), which are detected on the basis of receiving child and carer benefit payments, respectively.

3.4 Feature Engineering

We identify a spectrum of data elements characterizing an applicant's financial behavior next. Feature engineering is used to construct metrics which define and measure those behaviors to create a 'financial profile.' The transactions enable the construction of monthly inflows and outflows such as salary and benefits received, consumption patterns in relation to financial management ability (e.g., gambling habits, loan repayments, etc.) as well as temporal features such as volatility. We construct features across six proposed categories: (1) financial management, (2) financial distress, (3) financial resilience, (4) financial planning, (5) financial aid and (6) financial inclusion. Taken together, these enable applicant characterization based on level of financial health, degree of financial management ability and overall stability. The formulation of each category is detailed in the following Figure 2.

3.4.1 Financial Management

We define summary statistics of applicants' *inflow* behavior as the average monthly total income, average monthly salary and non-salary income, number of unique salary sources and consistency of salary inflow (monthly or weekly) of each applicant. Income is computed as all major inflows into the applicants' bank account, which includes, salary-based income, non-salary-based income, benefits received, pension received, and loans received, found under the categories:

Indicator	Feature Description	Data Type
Financial Management *(Inflow)*	- Total income (incl. salary, non-salary, benefits, pension)	Continuous
	- Salary-based income	Continuous
	- No. of unique salary sources	Integer
	- Consistency of salary inflow (monthly, weekly)	Binary
Financial Management *(Outflow)*	*Static spending behavior*	
	- Total, flexible and fixed expenditure	Continuous
	- Number of transactions	Integer
	- Average transaction amount	Continuous
	- Expenditure by category: cash, charity & donation, child & school, eating out & takeaways, fashion & beauty, fun & leisure, groceries & housekeeping, health & fitness, housing, medical & health, subscriptions, transport & fuel and utilities	Continuous
	Gambling behavior	
	- No. of gambling transactions	Integer
	- Gambling income and expenditure	Continuous
	Temporal spending behavior	
	- Persistence in fixed and flexible expenditure (bi-monthly and daily)	Continuous (0 to 1)
	- Persistence in categorical spending (monthly)	
	- Burstiness in fixed and flexible expenditure	Continuous (-1 to 1)
Financial Management *(Volatility)*	- Volatility in account balance, fixed & flexible expenditure, income and salary	Continuous (0 to 1)
Financial Distress	- Buy Now Pay Later financing (BNPL) usage	Continuous
	- Debt management (DM) & insolvency expenditure	Continuous
	Overdraft (OD) usage	
	- No. of days per month in OD	Integer
	- Proportion of months in OD (min. one day a month)	Continuous
	- Cost of OD fees	Continuous
	- Returned Direct Debits (RDD) payments and no. of transactions	Continuous
Financial Resilience *(Account Balance & Financial Shock Withstanding)*	- Account Balance: mean, minimum, and maximum value	Continuous
	- No. of transactions	Integer
	- Proportion of months one can withstand financial shock of £100	Continuous
	- Disposable income	Continuous
Financial Planning *(Insurance & Savings)*	- Holds any insurance and/or pension	Binary
	- Insurance and pension expenditure	Continuous
	- Holds voluntary insurance (e.g., appliance, life or pet)	Binary
	- Savings expenditure	Continuous
Financial Aid *(Benefits & Pension)*	- Total benefits	Continuous
	- Carer allowance	Continuous
	- Child benefits and tax credits	Continuous
	- Disability benefits	Continuous
	- Pension income	Continuous
Financial Inclusion *(Credit Card & Loan Usage)*	- No. of unique credit card providers (traditional and non-traditional)	Integer
	- Credit card payments	Continuous
	- No. of unique loan providers (traditional and non-traditional)	Integer
	- Loans received	Continuous
	- Loans paid	Continuous
	- Payday loan payments	Continuous

Figure 2: Financial management and financial difficulty features engineered from transaction data. *Note: All amount-based attributes are calculated as monthly averages across the applicant's account history in £/GBP unless noted otherwise.*

earnings and *credit bank transfers*. However, transactions considered as *internal transfers*, *returned direct debits* and *returns* are excluded from income calculation. To further distinguish between salary-based and non-salary income, transaction references containing select non-salary patterns (e.g., mobile transfers and gambling-related inflows) are excluded as well as transaction amounts less than £100 and multiples of £10, as these are majorly small money and/or mobile transfers made from other individuals or family members. Salary and non-salary income are differentiated because in the presence of liquidity constraints, another important resource individuals may consider are loans or gifts from family and friends, which can ensure minimum levels of consumption [Midões and Seré, 2022]. While these loans are typically short-term and small [Long, 2020], they do provide an additional financial buffer. Therefore, formal and informal credit options are both acknowledged to better assess FV. The number of unique salary sources is calculated using the references of salary-labeled transactions where unique references are counted as distinct sources. Salary inflows are labeled as consistent if more than 70% of an applicant's salary-based income is received in the same 10-day window every month (monthly) or three-day window every week (weekly).

We define summary statistics of applicants' *outflow* behavior as the average monthly total expenditure, including fixed and flexible expenditures (Figure 3), number of transactions, average transaction amount, and monthly expenditure by spending category including *cash, charity and donation, child and school, eating out and takeaways, fashion and beauty, fun and leisure, groceries and housekeeping, health and fitness, housing, medical and health, subscriptions, transport and fuel* and *utilities*. Given the association between gambling, addiction and financially harmful outcomes, gambling expenditures are considered separately. This includes metrics for average number of gambling transactions per month and average monthly gambling income and expenditure. To identify gambling transactions, Natural Language Processing (NLP) techniques of the transaction reference are used; further details can be found in the Appendix (Figure A.1).

Expenditure Types

Fixed		*Flexible*
bank transfer	housing	cash
child and school	insolvency provider	charity and donation
council tax and rates	insurance and pension	eating out and take aways
credit card payments	loans and repayments	fashion and beauty
debit internal transfers	medical and health	fun and leisure
debt enforcement and fines	payday loans	gambling
debt management	transport and fuel	groceries and housekeeping
fees and charges	tv licence	savings and investments
health and fitness	utilities	subscriptions

Figure 3: List of categories organized as fixed versus flexible expenditures.

Measures of persistence, burstiness and volatility are calculated next in order to consider temporal patterns. *Persistence* is used to evaluate the consistency in the amount an applicant spends in a monthly and weekly observation period $t' = \{M, W\}$; this metric is computed using the average cosine similarity coefficients between adjacent time intervals. For the monthly observation period, first bi-monthly spending (i.e., two elements for each month) are aggregated followed by the fraction of spending in each element. The persistence in spending amount is then calculated as the average of the cosine similarity:

$$persistence_{Monthly} = \frac{\Sigma_{i=0}^{n-1} cos(S_i, S_{i+1})}{n} \tag{1}$$

where S_i represents the vector of the relative amount spent in each bi-monthly (two-week) interval in a particular month i, and n represents the number of months we have of each applicant. This enables comparison between the first half of the month to the first half of the next month and subsequent months. A persistence value of zero implies that the relative amounts spent are dissimilar between the time intervals, whereas a value of one indicates they are the same across the time intervals. Similarly, weekly persistence is also calculated by grouping the spending amounts on a daily basis (i.e., 7 elements for each week). This enables comparison between consumption on a Monday one week and a Monday in subsequent weeks, for example. Persistence is calculated for both fixed and flexible expenditure amounts as well as the amount spent in each spending category; for example, to determine if an applicant consistently spends similar amounts in *child and school* items every month. These features are derived to gain a sense of diversity and stability within each spending category over time. However, because the value of transactions can be biased towards high-value categories (i.e., purchasing a washing machine), metrics based on the total number of transactions, as opposed to total value of the transactions, are also calculated to measure the frequency of purchasing activities across various categories.

Bursty dynamics, or *burstiness*, is defined as the heterogeneous property of time series which have short periods of intense high-frequency activities alternating with long periods of low-frequency activities [Tovanich et al., 2021]. This metric is used to measure the intensity of expenditure patterns. Burstiness is computed by first taking inter-event times, or the daily difference between two subsequent transactions. In our case, the transaction date is considered given time of purchase is not available. The inter-event time is defined as $t_i = T_i - T_{i-1}$ where T_i represents the transaction conducted at time i. The burstiness parameter is calculated as:

$$B = \frac{r-1}{r+1} \quad , \quad r = \frac{\sigma}{\mu} \tag{2}$$

where μ is the mean and σ is the standard deviation (SD) of the transactions' inter-event times. The burstiness parameter is calculated for all expenditure transactions, fixed and flexible, which reflects how regularly an applicant makes purchases daily. A burstiness value (B) of negative one indicates the purchasing pattern is completely stable, zero indicates random behavior and one indicates extreme and unpredicted spikes in expenditure behavior.

Lastly, we are also interested in the role of volatility, particularly regarding applicant's account balance, income, salary, and expenditure. This is inspired by the growing policy literature highlighting the importance of financial stability, and further intensified by recent phenomena such as 'zero-hour contracts,' the gig economy and major economic disruptions like the COVID-19 pandemic. The volatility parameter is computed by measuring the inter-month variation in average account balance (stock), income and salary (flow) and fixed and flexible expenditure amounts. The second-order coefficient variation [Kvålseth, 2017] is used to avoid issues caused by standard measures of variation which are sensitive to mean and outliers.

$$Volatility = \sqrt{\frac{(\frac{\sigma}{\mu})^2}{1 + (\frac{\sigma}{\mu})^2}} \tag{3}$$

where μ is mean and σ is SD over the full transaction history. Volatility is expressed as a value between zero and one, with zero indicating low volatility and one indicating high volatility.

3.4.2 Financial Distress

An applicant's ability to handle financial distress is measured by the following: amount of *debt management* (DM) and *insolvency* expenditure, usage of OD and having RDDs, all of which indicate insufficient funds in the account. The use of Buy Now Pay Later financing (BNPL) is also considered here. BNPL-financing programs (i.e., Klarna, Afterpay, Affirm), also known as point-of-sale loans, have recently become a popular layaway option enabling consumers to buy an item and then split the cost over a few weeks or months with regular installment payments. Usage of such programs may be indicative of situations where an individual does not have sufficient funds in their account to afford paying the full item's cost at once. However, it may also be the case that due to being typically interest-free credit, it is an attractive option for financially savvy users as well. OD usage is measured by the average number days per month in OD, proportion of months in consistent OD and fees paid. An account is considered in OD when the balance is a negative value for more than one day. RDDs, also known as bounced direct debits, occur when a bank rejects an online check or direct debit, typically when there are insufficient funds in the account to cover the amount requested. This metric is measured by counting the average number of RDD transactions per month.

3.4.3 Financial Resilience

Financial resilience refers to the ability of an individual to withstand unexpected life events that impact his or her income or assets. Financially stressful events can include unemployment, divorce, disability and medical problems. Therefore, to quantify an individual's financial resilience, features related to account balance value, disposable income and ability to withstand a financial shock of £100 are assessed. To determine an individual's 'assets,' the account balance value is used, with average monthly mean, minimum and maximum recorded. Disposable income which is assumed to be the net amount of money an individual has after paying necessary living expenses, including all fixed expenditures as well as *groceries and housekeeping* expenses (which were originally categorized by the data donor as flexible) is computed as: Income (excl. loan-based income) − all fixed expenditures − *groceries and housekeeping* expenditure. The proportion of months an individual can withstand financial shock is computed as the number of months an individual can withstand financial shock over the total number of months in their account history.

3.4.4 Financial Planning

Financial planning refers to the notion that an individual has a long-term financial strategy in place, represented by the use of insurance, pension, and savings. Holding insurance and consistently paying monthly premiums indicates the ability to protect oneself against unexpected costs and demonstrates financial responsibility. Holding more optional insurance options, such as pet or appliance insurance, may also indicate high financial responsibility. Measures include whether an individual holds any type of insurance and/or pension and the respective monthly average expenditure on each. The monthly average amount put towards savings is also computed based on the categories *debit internal transfers* and *savings & investments*.

3.4.5 Financial Aid

Financial aid refers to monetary assistance given to certain individuals on a conditional basis such as whether an individual is a carer, has a child, has a disability or is unemployed. As a result, they can receive monthly sums to help alleviate their expenses through carer allowance, child benefits and child tax credits, disability benefits, working tax credit, universal credit and employment support allowance respectively. More information regarding benefit and financial support provided by the UK government can be found on their website (https://www.gov.uk/ browse/benefits). We also include pension income in this category. For each type of benefit, the monthly average value received and its proportion to the individual's total income are computed.

3.4.6 Financial Inclusion

Financial inclusion refers to the ability of individuals to access useful and affordable financial products and services that meet their financial needs. We consider the use of credit cards and loans as part of this category. However, over-indebtedness is also a key indicator of FV therefore we observe the use of multiple credit or loan providers as evidence of this status. To compute the number of unique credit card providers used, individual providers are identified using NLP techniques, detailed in the Appendix (Figure A.2). Traditional providers refer to 'high street' banks such as Capital One, Barclays and HSBC whereas non-traditional providers refer to newer, challenger banks such as Starling and Monzo. This distinction is applied also to loan providers. The monthly average value of credit card payments, loans received, and loans paid are computed.

The aforementioned features are used directly as variables in the prediction and clustering models to generate the results shown in Section 4. A descriptive analysis can be found in the Appendix (Figures A.3, A.4) which displays the demographics of our data along target FV indicators and profile features. For example, a monotonic relationship can be seen across variables such as income and expenditure amounts which increase with age, corresponding to the greater number of financial responsibilities and spending that arise. However, later a dip is seen in applicants in the highest age ranges (ages 56 and over), most likely due to unemployment and eventually retirement.

4 Empirical Analysis and Results

Our aim is to understand to what extent individuals' FV can be inferred from transaction records, and how these are connected to protected characteristics. To first provide a comprehensive analysis of how one's financial behavior is associated with individual traits, we use the Pearson correlation coefficient to explore the association between the engineered features with FV indicators and protected characteristics. The complete results are available upon request; however, we briefly highlight select significant correlations (p-value < 0.001). Age is positively correlated to employment length, ability to withstand financial shock and being an owner occupier, likely due to greater stability in one's employment status thus ensuring more consistent money inflows which help withstand shock, and is negatively correlated with living with parents, likely due to financial freedom. Older individuals have greater total expenditure amounts (*groceries and housekeeping*, *housing* and *utilities* in particular) and higher persistency in categorical spending as well as less burstiness in flexible expenditures. Being female is positively correlated with having child-caring responsibilities and negatively correlated with gambling activities and living with parents, supported by the higher likelihood of having a child. Females tend to spend more in *fashion and beauty* and *groceries and housekeeping* categories along with greater use of BNPL financing options, which may coincide with their growing popularity in the online retail and clothing market [C+R Research, 2021], as well as higher non-traditional loan usage and number of overall transactions. Being a carer is positively correlated with having a child, likely due to the mediating impact of being female and is also highly correlated with having a disability, which may be due to a carer receiving disability benefits on behalf of the disabled individual for whom they are caring for. Spending on gambling activities is positively correlated with having higher disposable income, implying that those who gamble often gamble with funds remaining after paying for fixed expenses and not necessarily at the risk of going into OD (i.e., spending what they don't have). Gamblers tend to have higher burstiness in flexible expenditure, indicating that they are more likely to have extreme

and unpredicted spikes in expenditure behavior. Overall, we see that the correlations between FV indicators tend to be associated with one another, implying that exhibiting one indicator implies high likelihood of having others. Moreover, the correlations between the FV indicators found in our work reflect the findings of other FV studies [Daud et al., 2019, Financial Conduct Authority, 2015, O'Connor et al., 2019]. As a caveat, it is important to note that the identified associations are correlational as opposed to causation driven. Yet these associations are provided to help interpret the predictive models described in the following rather than being prescriptive in their own right.

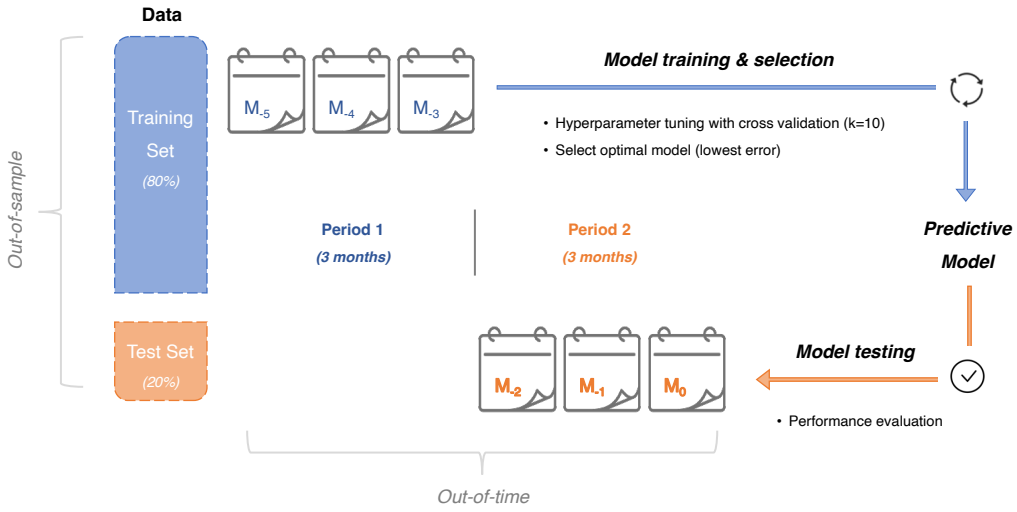

Figure 4: Data and empirical strategy.

Subsequently, the prediction task is devised as a binary classification problem (Figure 4). In light of exploding data volumes, ML serves as one of the most effective methods in data mining research [Dastile et al., 2020]. Therefore, three different classification algorithms are evaluated – Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Feature importance is subsequently calculated using the XGBoost model to determine the weightings of the predictive variables. For each model type, the dataset is randomly divided into 80% training set and 20% test set while retaining class ratios. Each set spans different three-month time periods thus ensuring out-of-time and out-of-sample testing. The events (e.g., performance along FV indicators) lead to imbalanced datasets where a minority of applicants typically have consistent financial trouble (frequencies are given in the earlier section on FVI). Therefore, to mitigate this imbalance, the majority class is randomly sampled to produce a balanced training dataset and the obtained model is then tested using realistic settings for the testing window. During the training phase, the model parameters are tuned using grid search with ten-fold cross-validation and tested against a 20% holdout set in the next time window shifted three-months forward. The classifier's performance is measured as the area under the receiver operating characteristic curve (AUROC) and the feature importances are estimated in terms of (normalized) relative influence. Other performance metrics are also recorded, including accuracy, precision and F1-score. The AUROC metric is focused on given its usefulness in classification scenarios where the trade-off between true positive rate and false positive rate is of vital interest.

4.1 Classification Performance

Figure 5 displays the predictive performance of LR, RF and XGBoost models for each FV indicator. The performance of the ML models is higher when classifying FV indicators as well as sociodemographic features. The highest performances are obtained with XGBoost to predict whether an applicant is an avid gambler (AUROC = 0.911) and whether they are unable to withstand financial shock (AUROC = 0.895) and the lowest performance when classifying insufficient disposable income (AUROC = 0.819) and insolvency (AUROC = 0.832). This may be due to the significant amount of time needed to reach insolvent status and thus requiring a debt management plan. Reaching this state likely implies the option to OD has already been cancelled by the bank and the individual has applied for payday lenders previously, with high amounts of loan payments or receipts. In other words, reaching insolvency status implies the individual was already financially vulnerable (months or years ago) therefore this indicator is likely not representative of their current financial state (i.e., last six months). Furthermore, predicting whether applicants exhibit protected characteristics is accomplished with decent performance, indicating that "blind" lenders may, in fact, be able to infer personal sociodemographic information even without direct access to it.

			Accuracy		Precision		Recall		F1 measure		AUROC	
			Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Financial Vulnerability Indicators	Financial shock withstanding	*LR*	0.693	0.006	0.693	0.012	0.693	0.007	0.693	0.008	0.756	0.007
		RF	0.818	0.005	0.818	0.008	0.818	0.010	0.818	0.005	0.903	0.004
		XGB	0.812	0.004	0.811	0.007	0.811	0.009	0.811	0.004	0.895	0.004
	Gambler	*LR*	0.811	0.005	0.718	0.016	0.796	0.015	0.740	0.016	0.875	0.006
		RF	0.806	0.004	0.718	0.011	0.806	0.013	0.739	0.014	0.893	0.006
		XGB	0.817	0.005	0.731	0.014	0.824	0.018	0.754	0.015	0.911	0.006
	Insolvent	*LR*	0.637	0.003	0.529	0.300	0.679	0.001	0.454	0.002	0.739	0.014
		RF	0.696	0.003	0.548	0.300	0.782	0.003	0.500	0.006	0.870	0.014
		XGB	0.697	0.003	0.544	0.192	0.752	0.006	0.495	0.012	0.832	0.015
	Insufficient disposable income	*LR*	0.603	0.004	0.555	0.000	0.648	0.000	0.497	0.000	0.706	0.014
		RF	0.706	0.003	0.578	0.373	0.686	0.002	0.564	0.003	0.754	0.013
		XGB	0.715	0.004	0.599	0.091	0.742	0.006	0.588	0.011	0.819	0.011
	Overdraft	*LR*	0.703	0.006	0.694	0.007	0.720	0.004	0.691	0.005	0.776	0.006
		RF	0.796	0.005	0.771	0.005	0.794	0.005	0.779	0.004	0.877	0.007
		XGB	0.796	0.005	0.773	0.003	0.799	0.005	0.780	0.003	0.881	0.005
	Returned direct debits	*LR*	0.693	0.003	0.678	0.008	0.699	0.012	0.677	0.011	0.775	0.005
		RF	0.763	0.004	0.753	0.009	0.785	0.011	0.753	0.009	0.870	0.004
		XGB	0.782	0.004	0.768	0.009	0.801	0.013	0.771	0.009	0.884	0.004
Benefits Received	Is a Carer	LR	0.861	0.002	0.564	0.060	0.811	0.017	0.577	0.022	0.895	0.028
		RF	0.860	0.002	0.573	0.057	0.869	0.017	0.589	0.026	0.945	0.013
		XGB	0.854	0.002	0.565	0.061	0.837	0.027	0.577	0.038	0.918	0.014
	Has a Child	LR	0.745	0.006	0.736	0.008	0.750	0.027	0.738	0.016	0.824	0.006
		RF	0.832	0.004	0.822	0.007	0.840	0.007	0.826	0.005	0.917	0.003
		XGB	0.831	0.005	0.821	0.009	0.838	0.007	0.825	0.006	0.917	0.004
	Has a Disability	LR	0.743	0.003	0.581	0.041	0.718	0.011	0.575	0.016	0.796	0.018
		RF	0.774	0.004	0.613	0.048	0.801	0.012	0.621	0.017	0.885	0.010
		XGB	0.766	0.004	0.605	0.050	0.779	0.012	0.609	0.018	0.864	0.010
	Female	LR	0.719	0.005	0.699	0.005	0.756	0.004	0.694	0.004	0.832	0.004
		RF	0.779	0.005	0.746	0.006	0.810	0.003	0.753	0.003	0.896	0.003
		XGB	0.815	0.005	0.774	0.006	0.836	0.003	0.788	0.004	0.917	0.002

Figure 5: Classification models' performance. *Note: ML models' performance (LR = logistic regression, RF = random forest, XGB = XGBoost) evaluated along accuracy, F1 score, precision, recall and AUROC. Real cut-off value is not known and may differ between different lenders therefore AUROC is used for interpretation.*

4.2 Feature Importance

To provide interpretability of the outcome, an investigation into the importance of features used by the model is conducted. This enables us to gain insight into which features are most relevant to FV and thereby useful for identifying potential risk of indirect discrimination or proxies. The results are presented in Figure 6, showing the top eight features for each target. We rank all features based on their magnitude of importance using the XGBoost model given their high performance. The importance is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.

Supporting the correlation results, the most significant features for predicting FV indicators are typically other indicators, indicating strong linkages between the metrics. For example, predicting whether an individual can withstand financial shock puts importance on features such account balance volatility, savings, number of RDDs and having been insolvent. Predicting insolvency heavily weighs days in OD and number of traditional loans while predicting OD similarly puts importance on RDDs, savings, and volatility in account balance and income. To predict gambling, burstiness in flexible expenditure and gender are important features which is supported by the correlation analysis. Having a child is heavily based on being female, living with parents and *child and school* spending; being a carer on disability benefits, salary and its volatility; and having a disability on carer benefits, likely due to the overlap between those disabled receiving disability benefits directly and those who are carers of disabled individuals receiving benefits on their behalf. Due to the ambiguity of the transaction labels and redacted references, the distinction cannot be further refined.

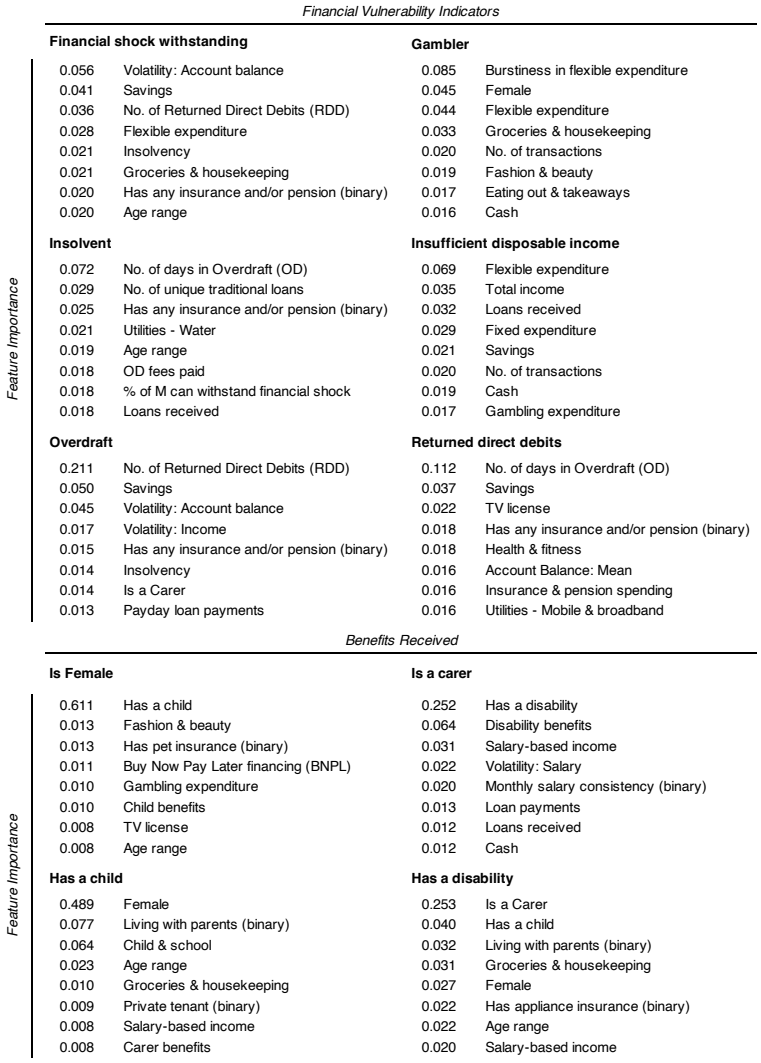

Figure 6: Interpretability of classification models' performance using feature importance, which is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.

4.3 Clustering Approach

Clustering is used in our analysis for the task of demonstrating how segmentation can implicitly capture sensitive and protected characteristics, even when they are not included into the analysis. There are multiple applications of clustering in financial data analytics, particularly in behavior analysis [Thompson et al., 2020] and marketing, such as Customer Relationship Management (CRM) strategies [Roshan and Afsharinezhad, 2017] which segment and profile customers according to their needs, desires, or distinguishing characteristics such as age, ethnicity, profession, gender and location or psychographic factors such as shopping behavior, interests, and motivation [Hsieh, 2004]. With this analysis, we highlight the need for lenders to be concerned not only about the association of certain financial behaviors with a particular sensitive or protected characteristic, but also with the *combination* of several protected characteristics.

RFM (Recency, Frequency, Monetary) is a widely used method to quantitatively analyze customer behavior based on three dimensions: how recently customers buy, how often they buy, and how much they spend [Cheng and Chen, 2009, Wei et al., 2010]. By understanding the history, number and value of customers' transaction updates, this method can help identify those more likely to respond to certain promotions or for enhancing personalized services. When applied to credit providers, "better" service can include customized loan options, such as loan size, conditions of repayment, or variable interest rates best suited to each applicant's financial profile. Similarly, we are interested in identifying common profiles of applicants representing different degrees of financial health and financial management ability within

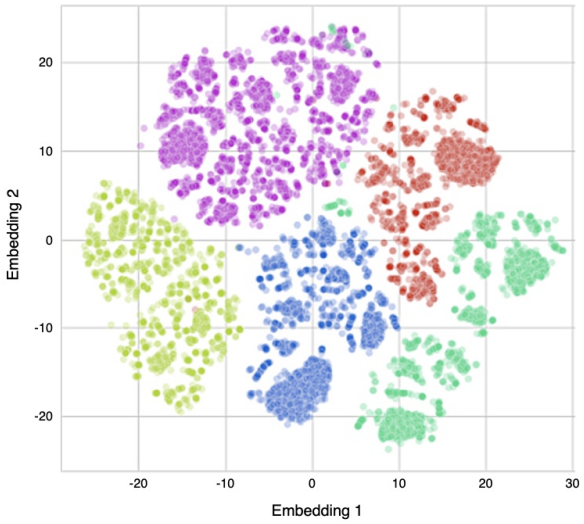

Figure 7: t-SNE visualization for the full dataset by cluster projected onto two embeddings.

our applicant pool. This can be used to determine a range of affordability scenarios for different types of applicants and serve as contextual information for improved affordability assessment measures by lenders.

Principal component analysis is first used to reduce the dimensionality of the data into 15 principal components representing 78.1% of the cumulative variance. This is followed by the k-means algorithm. To find the optimal number of clusters (k), two assessment measures are applied – the elbow method and silhouette approach – the inflection point and average value maximization respectively point to the use of five clusters. Finally, t-Distributed Stochastic Neighbor Embeddings (t-SNEs) are used for the purpose of visualizing the high-dimensional data. By projecting high dimensional data onto a lower-dimensional space, known as embeddings, t-SNE preserves the local data structure [Maaten and Hinton, 2008] and forms a non-linear mapping, thus keeping similar data points (i.e., applicants) closer together in the low-dimensional space for visualizing clusters. A perplexity of 300 is used to obtain a stable embedded plot (Figure 7). The data cleaning, feature engineering, correlation, model prediction, clustering algorithm, t-SNE visualization, and analysis are implemented using Python version 3.6. The t-SNE algorithm used for data visualization can be found in the sklearn Python package [Pedregosa et al., 2011]. From the two-dimensional embedding map, we can see that there are distinct boundaries between most of the clusters with overlaps between clusters 4 and 5 (green and purple, respectively), however, it is worth noting that higher dimensional embeddings can reveal other higher-order boundaries that may distinguish overlapping clusters. Therefore, the projection from three dimensions to two dimensions, such as in this case for visualization purposes, may create the appearance of overlap.

To facilitate the interpretation of the clustering results, the mean values of features of interest (including those that were not used in obtaining clusters, such as demographics) for each cluster are displayed in Figure 8. We first highlight key patterns and defining characteristics to understand the behaviors shared. Clusters 1 and 3 are similar in their credit usage however significantly differ in terms of inflow (income, salary) and outflow (fixed, flexible expenditure). Clusters 2 and 3 are similar in their demographics and income levels, but cluster 3 spends significantly more in expenditures and have lower volatility overall. Cluster 4 is unique with the lowest age and lowest amount of income and salary and most notably, cluster 5 holds a significantly high percentage of female applicants and those with a child. A high-level overview is visualized as a heatmap along key differentiating features, including FV indicators and select inflow and outflow metrics (Figure 9). We note that cluster 3 performs the best along all FV indicators, implying they are the most financially healthy. Next, we summarize in greater detail the unique, distinguishing attributes of each cluster.

4.4 Applicant Profiles

While individuals from different groups may appear similar, they are classified based on subtle differentiating factors determined by the clustering algorithm. Therefore, we narrate "profiles" of applicants to ease discussions and better understand the clusters as real people, not simply data points. *Cluster 1 – The Credit User* (19.4% of applicants): These applicants use significant amounts of financial aid and credit options in the form of loans, the use of a payday lender, credit cards and BNPL financing. Relative to other clusters, this group has the highest credit card usage and payments, across both traditional and non-traditional providers in addition to the highest loan usage, both traditional

		Cluster				
		1	2	3	4	5
Profile	Age range	3.4	4.4	4.4	3.1	3.6
	Female	0.7	0.5	0.8	0.6	1.0
	Housing association tenant (binary)	0.2	0.2	0.2	0.2	0.3
	Living with parents (binary)	0.3	0.2	0.1	0.3	0.1
	Owner occupier (binary)	0.1	0.2	0.2	0.1	0.1
Financial Management	Total income	2867	2942	3338	2490	3008
	Salary-based income	1587	1651	1825	1314	1298
	Fixed expenditure	2281	2198	2289	1846	2296
	Flexible expenditure	1263	1329	1808	1145	1463
	Child & school	4.5	3.3	9.9	2.3	15.7
	Eating out & takeaways	76.9	57.0	91.7	66.7	84.1
	Fashion & beauty	95.5	64.9	119.8	60.4	108.7
	Groceries & housekeeping	330.1	292.7	536.2	229.5	448.2
	Housing	84.1	99.0	132.8	60.4	113.9
	Fun & leisure	50.6	41.2	70.1	35.9	56.8
	Medical & health	5.4	4.2	8.0	3.3	5.3
	Transport & fuel	83.7	74.9	83.8	53.8	74.7
	Utilities total	187.8	174.0	239.9	124.5	223.5
	Gambling expenditure	134.4	233.3	254.6	205.8	135.2
	Volatility: Account balance	0.5	0.5	0.3	0.5	0.5
	Volatility: Fixed expenditure	0.4	0.4	0.4	0.4	0.4
	Volatility: Flexible expenditure	0.4	0.5	0.4	0.5	0.4
	Volatility: Income	0.4	0.4	0.3	0.4	0.3
	Volatility: Salary	0.3	0.3	0.3	0.4	0.4
Distress	Buy Now Pay Later financing (BNPL)	60.9	0.0	56.2	24.4	52.2
	Insolvency & debt management (DM)	29.6	28.2	39.2	17.3	27.7
	No. of days in Overdraft (OD)	8.1	7.7	0.9	6.8	7.7
	No. of Returned Direct Debits (RDD)	1.1	0.9	0.4	0.5	1.1
Resilience	Account Balance: Mean	74.2	103.1	662.4	51.5	88.5
	% of M can withstand financial shock	0.4	0.4	1.0	0.4	0.4
	Disposable income	748.7	719.8	878.3	703.6	824.4
Planning	Insurance & pension spending	66.4	63.6	54.6	0.0	50.1
	Savings	330.9	269.5	136.4	326.7	418.8
Aid	Total benefits	110.3	94.9	378.1	129.7	677.3
	Carer benefits	1.8	2.3	7.3	2.6	8.2
	Child benefits	0.0	0.4	94.6	1.2	187.1
	Disability benefits	16.4	16.3	47.1	17.3	40.9
	Pension income	5.8	11.8	7.8	4.2	2.4
Inclusion	CC payments	44.9	42.5	44.5	25.9	26.3
	No. of unique traditional loans	0.9	0.7	0.8	0.5	0.6
	No. of unique non-traditional loans	3.5	0.9	2.7	1.9	2.6
	Payday loan payments	69.5	60.8	48.9	55.1	44.7

Figure 8: Mean values of the profile and financial management and difficulty features for each cluster, including the features that were not used in obtaining the clusters. *Note: For each feature, lowest values are highlighted in red and highest values are highlighted in green.*

and non-traditional, particularly with a payday lender. The amount of loans paid makes up 16% of their income, the highest proportion relative to other clusters. These applicants have a high volatility in average account balance, highest proportion of days in OD and highest number of RDDs. However, they are less likely to have dependents with minimal disability or carer benefits as well, and are more likely to live with their parents, thereby reducing their housing costs.

Cluster 2 – The Financially Resilient (15.9% of applicants): These applicants represent the second oldest age group with the highest insurance and pension spending but with low to no benefits received. They have a low likelihood of having a child, however, receive the second highest amount of pension compared to cluster 3, coinciding with their average age range. This group has low average disposable income along with low amounts of voluntary expenditures in categories such as *eating out and takeaway* as well as *fashion and beauty*. They also hold a low likelihood of non-traditional loan usage, however, have high usage of a payday lender.

Cluster 3 – The Financially Secure (19.0% of applicants): These applicants represent the most financially secure and healthy group of individuals with the highest inflows and highest outflows. They represent the oldest age group which correlates with their higher employment length and may explain the financial security evidenced by having the highest average account balance and disposable income relative to other clusters. They are also able to withstand financial shock nearly 100% of the time. This group is characterized by a consistent monthly salary, lowest number of days in OD

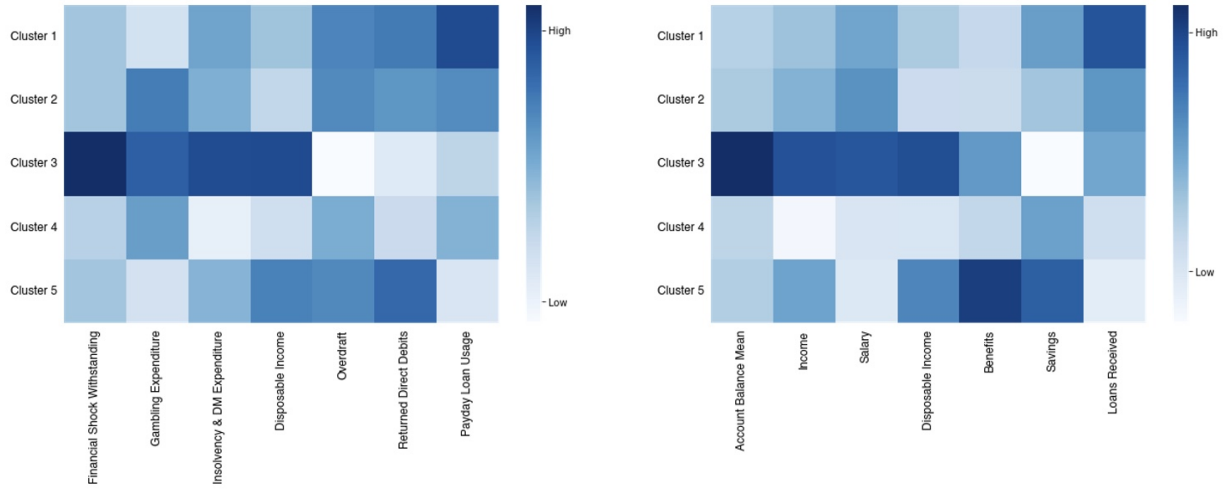

Figure 9: Heatmap display of the key select differentiating features across the five clusters.

and lowest volatility in account balance, income, and salary. They also receive a high proportion of disability benefits and pension-based income. Alongside higher inflows, this group also displays high expenditure amounts, particularly with the highest flexible expenditure amounts across all spending categories (e.g., *eating out, fashion, fun and leisure, subscriptions, charity, housing, medical and health, groceries, gambling* and *utilities*). This group holds the largest proportion of owner occupiers which coincides with their financial security.

Cluster 4 – The Young and Challenged (16.7% of applicants): These applicants represent the youngest group with high volatility in salary and income but correspondingly with the lowest expenditures, both fixed and flexible. This cluster has the lowest earnings relative to other groups but also the lowest spending with high burstiness in flexible spending, indicating uncertainty and a more volatile lifestyle. They are unable to withstand financial shock for most parts of the year (65% of months) along with their low average account balance of £51. Their salary represents a smaller proportion of their overall income with higher likelihood of weekly salary inflows (as opposed to the monthly standard), implying alternative income sources. This group receives minimal benefits, most likely due to their young age, with the lowest usage of traditional loans but higher usage of non-traditional loans, which may reflect younger adults shifting towards FinTech products. This group has the highest likelihood of living with their parents and lowest of being an owner occupier, which accounts for their significantly low *utilities* expenditure, particularly in energy usage.

Cluster 5 – The Beneficiary (29.1% of applicants): These applicants hold the highest proportion of females relative to other clusters alongside having the highest likelihood of holding responsibilities such as having a child or being a carer. As a result, this group subsidizes their living costs with a significant proportion of benefits, making up 28% of their overall income. Their categorical spending skews largely towards *child and school* expenditure while saving *housing* costs by being a housing association and/or council tenant. They have minimal *gambling* expenditure, likely due to the higher proportion of females. This group is also characterized with the highest number of RDDs with a high proportion of days spend in OD and high BNPL usage. However, they are not as likely to use loans, a payday lender or receive pensions. Interestingly, they are characterized with the highest volatility in salary but lowest volatility in income relatively, implying significant non-salary income support, mainly in the form of benefits. Therefore, when salaries are impacted for this demographic, they are still able to accrue consistent income from other public transfers.

Let us consider in greater detail clusters 3 and 5 (Figure 10), comparing two distinct groups in terms of socio-demographic profile and financial behavior. We can see that while salary inflows differ significantly, both clusters have comparable disposable income. This can be attributed by the significant proportion of benefits received by cluster 5 which compensates for their lack of salary, thus highlighting the importance of child and unemployment benefits to their financial wellbeing. With benefits making up 28% of their income on average, these individuals would be unable to cope with necessary life expenses without such support or if they are affected by the loss of supplementary income. This poses the question of whether lenders should be taking benefits and proportion of these benefits to overall income into consideration within their affordability criteria when assessing applicants. And if not, whether this may adversely affect women with children and thus risk indirectly discriminating against this group of individuals.

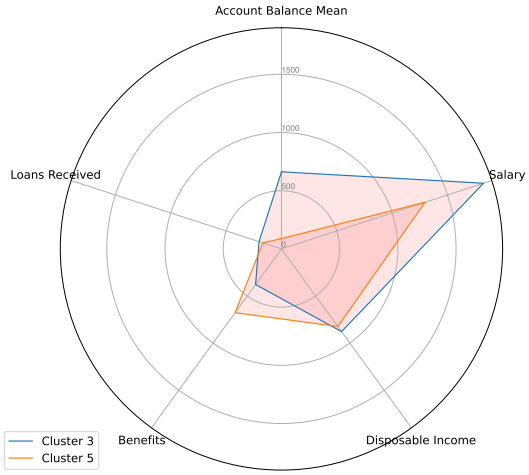

Figure 10: Radar chat comparison between cluster 3 *(The Financially Secure)* and cluster 5 *(The Beneficiary)*.

5 Discussion

Our study has involved a variety of approaches analyzing an Open Banking dataset to explore financial behaviors and their implications on FV and fairness. We propose a methodology for identifying and assessing FV using transaction data, including six major indicators for differentiating levels of financial stability, capacity and management ability. The engineered features can be used to improve the performance of predictive models regarding individualized credit risk, with better insight into FV warning signs particularly for at-risk populations. This may be useful for alternative lenders looking to expand their customer base into underbanked populations. The results suggest that transaction data can be highly predictive of future FV, three months ahead of model application. The kind of analysis described here can be used to judge new loan applicants for creditworthiness, where the given models are applicable to both applicants who can provide traditional credit data enhanced with transaction history (hybrid model) as well as applicants with limited or no credit scoring history (transaction-only model).

We've shown that protected characteristics can also be inferred, even when removed from the underlying data. This renders the fairness via mandated "blindness" approach (also known as "fairness through unawareness" [Dwork et al., 2011], which naively ignores all protected attributes, as futile in the current data environment. We would like to point out the risk of certain behavioral attributes serving as proxies for these attributes, enabling the prediction of omitted attributes through other permitted features. Lastly, the clustering results reveal two major implications. Firstly, they illustrate that decision-makers should remain concerned about the combination of protected characteristics with certain financial behaviors. Secondly, these results can help policymakers and practitioners form a profile of the type of consumers who are most at risk of "sliding" from a state of low to high FV over time and in need of the most support. For example, signaling one's financial health can be done incrementally before payment deadlines to inform the user or lending institution if preemptive remedies are needed before missing payments or defaulting. Doing so also helps address recent calls urging stakeholder groups to pay greater attention to the characteristics of consumers "trending toward vulnerability" [O'Connor et al., 2019, p. 427]. Recent household surveys reveal that one in five households have depleted their savings, fallen behind on housing payments, or are experiencing difficulty paying their debts, buying groceries, and paying utilities as a result of the COVID-19 pandemic [Center on Budget and Policy Priorities, 2020], stressing the timeliness and gravity of this issue. In the future, we would like to explore how decision making can be personalized for different applicant clusters. This has important implications for individuals as opportunities for data sharing in return for opportunities (i.e., a loan) grow, and organizations look to draw more detailed insights from their customers for marketing, retention or risk management purposes.

Some unique highlights are also mentioned. Firstly, our data set encompasses a particular demographic in the UK, who majorly work in the public sector and tend to be more financially challenged than average. This is not only evidenced in the data (i.e., salary and income amounts) but via selection bias of the applicants received. The dataset was collected from applicants who voluntarily applied to an alternative lender which differentiates itself by not requiring a standard credit check, thus attracting a niche demographic. While this may be seen as a limitation, we view it as an

opportunity to shed light on individuals often lost at tail ends of the general population; or those holding a different distribution over select features and thus a different relationship with the predicted label relative to others. Typically, models fitted to the majority which look to minimize overall error can result in representation bias – a key cause of unfairness [Mehrabi et al., 2019]. Secondly, there are limited studies that utilize standalone Open Banking data. Researchers have shown improved predictive capability by supplementing traditional credit models with alternative data [Djeundje et al., 2021, Gambacorta et al., 2019, Óskarsdóttir et al., 2019], however, there are no studies in the context of FV with consideration of fairness, to the best of our knowledge. This is also the first that provides insight tailored around an at-risk demographic. However, we believe our analytical methods can be applied to other populations with transaction-based datasets and hope our use of routine and temporal patterns for characterizing financial behavior will inspire future research.

We note that this study contains some limitations that could be addressed with future research. First, we do not establish causality, rather the findings demonstrate associations that may reflect causality or comorbidity – both of which are of concern. Causality would indicate, for example, higher levels of gambling increase one's risk of FV. Comorbidity, however, would indicate that individuals who are susceptible to such negative outcomes due to alternative factors are more likely to be drawn to gambling; for example, where having a bounced check leads to gambling as a means to pay off debt. In reality, the observed effects are likely a blend of both effects. Furthermore, while OD is considered a FV indicator in this study, it is worth noting the possibility of individuals without OD may, in fact, be more financially vulnerable because they do not have a facility to use in the case of financial shocks or alternatively do not have the option of OD to begin with because their bank considers them high risk. Further work is needed to measure the extent to which FV is driven by causal mechanisms. Nonetheless, this longitudinal transaction-based approach informs the current FV debate. Vulnerability characterization using financial behavioral markers, covering consumption, savings, and monetary inflow can be insightful for economic research and public policy to help identify individuals who may be more sensitive to income shocks; for example, distinguishing those paid on a consistent monthly basis versus those self-employed or largely living off benefits. It also informs how much households would have to reduce flexible expenditures or savings to maintain basic, necessary consumption levels. In that sense, this work adds to the existing literature on the use of Open Banking data as a means for assessing and possibly preventing FV.

5.1 Future Work and Applications

Our work ultimately looks to help promote financial sustainability. Naturally, this study is followed by asking what level of shock each consumer type can withstand and formulating strategies to increase the financial prospects of select groups. Future work looks to explore stress testing methods to discern how robust users are to certain types and levels of financial shock – a critical concern of credit providers but with applications to other financial institutions such as mortgage and insurance providers [Marron, 2007, Mester, 1997]. For example, a stable, responsible individual may be able to overcome a financial shock successfully with little, temporary help as opposed to a shopaholic or avid gambler who may be tempted to spend funds irresponsibly into OD. Understanding what form of credit would be beneficial for an individual, at what point in time, would be informative to lenders. Due to the recency of the applications, we are unable to discern repayment behavior therefore we plan to examine the linkages between FV indicators and repayment habits, while looking for evidence on whether we can nudge any of the noted behaviors.

In real-world settings, this information can be used to determine whether the demand for credit and capacity of those in need match the supply of options available in the current financing market. If it is not possible for this demographic to sustain the market rates, this begs the question of whether regulatory agencies should act to ensure more affordable lenders are available or publicly sponsored. Furthermore, with the appropriate checks and balances established, the observations presented here could be used to provide individual feedback and nudge healthier decision making. For example, financial support can be released in flexible ways, such as frequent smaller payments rather than a lump sum with real-time feedback provided via transactional data. If a consumer acts reasonably, the lender can decide to release additional funds or relax interest rates, enabling disadvantaged individuals to rise out of their debt cycle. This can act as a "financial rehab" of sorts for those deemed higher risk along traditional credit metrics but desire an opportunity to improve their financial standing over time, rather than remain limited to unaffordable or fringe lenders.

While all the data used in this study is anonymized and at no point are individuals identifiable in its undertaking, we remain highly cognizant of the implications surrounding behavior-to-outcome associations. As the data landscape grows more complex, data points are no longer singular but can have multiple contextual implications, fueling concerns over fairness, discrimination as well as privacy. We highlight that certain groups of individuals may be unknowingly or unintentionally disadvantaged even when seemingly 'neutral' data is utilized. In most cases, individuals will not expect or even be aware of the fact that their consumption data may reveal personal factors. Realistically even the most motivated users will find it increasingly difficult to accrue the knowledge required to make self-interested decisions which trade-off against immediate and tangible benefits that may come with data sharing. Therefore, we conclude by

challenging the claim that all algorithms are neutral and caution against an "all data is credit data" approach [Aitken, 2017]. As the potential for discriminatory harms magnify in proportion to the technological advancements, we provide a word of caution to lenders using Open Banking data where the risk of abuse can easily outweigh its potential benefits without proper oversight.

Acknowledgements The authors are grateful to the data provider who has chosen to remain anonymous. S. Kim is grateful to Baillie Gifford for funding.

References

Anindya Ghose, Xitong Guo, Beibei Li, and Yuanyuan Dang. Empowering patients using smart mobile health platforms: Evidence of a randomized field experiment. *MIS Quarterly*, 46:151–192, 3 2022. URL https://misq.umn.edu/empowering-patients-using-smart-mobile-health-platforms-evidence-of-a-randomized-field-experiment.html.

Xuechen Luo, Ling Ge, and Chong (Alex) Wang. Crowdfunding for microfinance institutions: The new hope? *MIS Quarterly*, 46:373–400, 3 2022. ISSN 02767783. doi:10.25300/MISQ/2022/15406. URL https://misq.umn.edu/crowdfunding-for-microfinance-institutions-the-new-hope.html.

Nydia Remolina. Open banking: Regulatory challenges for a new form of financial intermediation in a data-driven world. *SMU Centre for AI Data Governance Research*, 5, 2019.

Eva Wolkowitz and Sarah Parker. Big data, big potential: Harnessing data technology for the underserved market. *Center for Financial Services Innovation*, 2015.

Janine S. Hiller and Lindsay Sain Jones. Who's keeping score?: Oversight of changing consumer credit infrastructure. *American Business Law Journal*, 59:61–121, 3 2022.

R. K.E. Bellamy, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, and S. Mehta. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63, 7 2019.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv*, 2018.

Stefan Lessmann, Bart Baesens, Hsin Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247:124–136, 2015.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54, 8 2019.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. volume Part F1296, pages 797–806, 2017.

Rob Aitken. 'all data is credit data': Constituting the unbanked. *Competition Change*, 21:274–300, 2017.

Financial Conduct Authority. Financial lives 2020 survey: The impact of coronavirus, 2021. URL https://www.fca.org.uk/publications/research/financial-lives-2020-survey-impact-coronavirus.

Breno Braga, Signe-Mary Mckernan, William J Congdon, Elizabeth Mandiville, Mark Hayward, and Christopher Trepel. Financial vulnerability in the united kingdom from brexit to the covid-19 pandemic. *Lowell's European Consumer Research Laboratory*, 2021.

Carol Bruce, Maeve E. Gearing, Jill DeMatteis, Kerry Levin, Timothy Mulcahy, Jocelyn Newsome, and Jonathan Wivagg. Financial vulnerability and the impact of covid-19 on american households. *PLoS ONE*, 17, 1 2022. ISSN 19326203. doi:10.1371/JOURNAL.PONE.0262301. URL /pmc/articles/PMC8759691//pmc/articles/PMC8759691/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8759691/.

Coulter Jones, Jean Eaglesham, and AnnaMaria Andriotis. How payday lenders target consumers hurt by coronavirus. *The Wall Street Journal*, 2020. URL https://www.wsj.com/articles/how-payday-lenders-target-consumers-hurt-by-coronavirus-11591176601.

Financial Conduct Authority. Occasional paper no. 8: Consumer vulnerability, 2015. URL https://www.fca.org.uk/publications/occasional-papers/occasional-paper-no-8-consumer-vulnerability.

Kevin O'Leary, Philip O'Reilly, Tadhg Nagle, Christos Filelis-Papadopoulos, and Milad Dehghani. The sustainable value of open banking: Insights from an open data lens. Hawaii International Conference on System Sciences, 2021.

Nadine Côrte-Real, Tiago Oliveira, and Pedro Ruivo. Assessing business value of big data analytics in european firms. *Journal of Business Research*, 70:379–390, 1 2017.

Annette Vissing-Jorgensen. Consumer credit: Learning your customer's default risk from what (s)he buys. *SSRN Electronic Journal*, 2012.

Tao Zhang, Wei Zhang, Xu Wei, and Haijing Hao. Multiple instance learning for credit risk assessment with transaction data. *Knowledge-Based Systems*, 161:65–77, 2018.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34:2767–2787, 11 2010.

Jung Youn Lee, Joonhyuk Yang, and Eric Anderson. Buying and payment habits: Using grocery data to predict credit card payments. *SSRN Electronic Journal*, 6 2021.

Marion Fourcade and Kieran Healy. Seeing like a market. *Socio-Economic Review*, 15:9–29, 1 2017.

Marie Langevin. Big data for (not so) small loans: Technological infrastructures and the massification of fringe finance. *Review of International Political Economy*, 26:790–814, 9 2019.

Alessandro Acquisti, Leslie K. John, and George Loewenstein. What is privacy worth? *Journal of Legal Studies*, 42: 249–274, 7 2013.

Patricia A. Norberg, Daniel R. Horne, and David A. Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41:100–126, 6 2007.

Nizan G Packin. On social credit and the right to be unnetworked. *SSRN Electronic Journal*, 2016.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data and Society*, 4, 12 2017.

M. Hurley and Julius Adebayo. Credit scoring in the era of big data. *Yale Journal of Law and Technology*, 18, 2017.

Irene Y. Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 2018-December:3539–3550, 5 2018.

Michael Rovatsos, Brent Mittelstadt, and Ansgar Koene. Landscape summary: Bias in algorithmic decision-making: What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?, 2019.

Mark H. Lopez, Lee Rainie, and Abby Budiman. Financial and health impacts of covid-19 vary widely by race and ethnicity. *Pew Research Center*, 5 2020. URL https://www.pewresearch.org/fact-tank/2020/05/05/financial-and-health-impacts-of-covid-19-vary-widely-by-race-and-ethnicity/.

Pasqualina Perrig-Chiello, Sara Hutchison, and Bina Knöpfli. Vulnerability following a critical life event: Temporary crisis or chronic distress? a psychological controversy, methodological considerations, and empirical evidence. *Life Course Research and Social Policies*, 3:87–111, 2016.

S. Costa and L. Farinha. Households' indebtedness: A microeconomic analysis based on the results of the households' financial and consumption survey. *Financial Stability Report. Banco de Portugal*, 2012.

Young Il, Hyoung Chan, Joo Hee Article, By IL Young Kim, Hyoung Chan Kim, and Joo Hee Yoo. Household over-indebtedness and financial vulnerability in korea: Evidence from credit bureau data. *KDI Journal of Economic Policy*, 38:53–77, 8 2016.

Miguel Ampudia, Has van Vlokhoven, and Dawid Żochowski. Financial fragility of euro area households. *Journal of Financial Stability*, 27:250–262, 12 2016.

Sarah Bridges and Richard Disney. Use of credit and arrears on debt among low-income families in the united kingdom. *Fiscal Studies*, 25:1–25, 3 2004.

Andrew C. Worthington. Debt as a source of financial stress in australian households. *International Journal of Consumer Studies*, 30:2–15, 1 2006.

Annamaria Lusardi, Daniel Schneider, and Peter Tufano. Financially fragile households: Evidence and implications. *NBER Working Paper 17072*, 2011.

Financial Conduct Authority. Understanding the financial lives of uk adults: Findings from the fca's financial lives survey 2017, 2017.

Consumer Financial Protection Bureau. Measuring financial well-being: A guide to using the cfpb financial well-being scale. 2015.

Aldo Rustichini, Colin G. DeYoung, Jon E. Anderson, and Stephen V. Burks. Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation. *J Behav Exp Econ*, 64: 122–137, 10 2016.

Joe J. Gladstone, Sandra C. Matz, and Alain Lemaire. Can psychological traits be inferred from spending? evidence from transaction data. *Psychol Sci*, 30:1087–1096, 7 2019.

John Gathergood. Self-control, financial literacy and consumer over-indebtedness. *Journal of Economic Psychology*, 33:590–602, 6 2012.

Annamaria Lusardi. Financial literacy and the need for financial education: Evidence and implications. *Swiss Journal of Economics and Statistics*, 155:1–8, 12 2019.

William B. DeHart, Jonathan E. Friedel, Jean M. Lown, and Amy L. Odum. The effects of financial education on impulsive decision making. *PLoS ONE*, 11, 7 2016.

Kenneth P Brevoort, Philipp Grimm, and Michelle Kambara. Credit invisibles and the unscored. *Cityscape*, 18, 2016.

Financial Conduct Authority. Conc 1.3 guidance on financial difficulties - fca handbook. URL https://www.handbook.fca.org.uk/handbook/CONC/1/3.html.

Drew M. Anderson, Alexander Strand, and J. Michael Collins. The impact of electronic payments for vulnerable consumers: Evidence from social security. *Journal of Consumer Affairs*, 52:35–60, 3 2018.

Merlyn A. Griffiths and Tracy R. Harmon. Aging consumer vulnerabilities influencing factors of acquiescence to informed consent. *Journal of Consumer Affairs*, 45:445–466, 9 2011.

Research frontiers on older consumers' vulnerability. *Journal of Consumer Affairs*, 45:467–491, 9 2011.

Catarina Midões and Mateo Seré. Living with reduced income: An analysis of household financial vulnerability under covid-19. *Social Indicators Research*, 161:125, 5 2022.

Melanie G. Long. Informal borrowers and financial exclusion: The invisible unbanked at the intersections of race and gender:. *The Review of Black Political Economy*, 47:363–403, 7 2020.

Natkamon Tovanich, Simone Centellegher, Nacéra Bennacer Seghouani, Joe Gladstone, Sandra Matz, and Bruno Lepri. Inferring psychological traits from spending categories and dynamic consumption patterns. *EPJ Data Science*, 10, 12 2021.

Tarald O. Kvålseth. Coefficient of variation: The second-order alternative. *Journal of Applied Statistics*, 44:402–415, 2 2017.

C+R Research. Buy now, pay later statistics and user habits. 2021. URL https://www.crresearch.com/blog/buy_now_pay_later_statistics.

Siti Nurazira Mohd Daud, Ainulashikin Marzuki, Nursilah Ahmad, and Zurina Kefeli. Financial vulnerability and its determinants: Survey evidence from malaysian households. *Emerging Markets Finance and Trade*, 55:1991–2003, 7 2019.

Genevieve E. O'Connor, Casey E. Newmeyer, Nancy Yee Ching Wong, Julia B. Bayuk, Laurel A. Cook, Yuliya Komarova, Cazilla Loibl, L. Lin Ong, and Dee Warmath. Conceptualizing the multiple dimensions of consumer financial vulnerability. *Journal of Business Research*, 100:421–430, 7 2019.

Xolani Dastile, Turgay Celik, and Moshe Potsane. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263, 6 2020.

John R. J. Thompson, Longlong Feng, R. Mark Reesor, and Chuck Grace. Know your clients' behaviours: A cluster analysis of financial transactions. *Journal of Risk and Financial Management*, 14:50, 5 2020.

Hadi Roshan and Masoumeh Afsharinezhad. The new approach in market segmentation by using rfm model. *Journal of Applied Research on Industrial Engineering*, 4, 2017.

Nan Chen Hsieh. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, 27:623–633, 11 2004.

Ching Hsue Cheng and You Shyang Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert Systems with Applications*, 36:4176–4184, 4 2009.

Jo-Ting Wei, Shih-Yen Lin, and Hsin-Hung Wu. A review of the application of rfm model. *African Journal of Business Management*, 4:4199–4206, 2010.

Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9: 2579–2605, 2008.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL http://scikit-learn.sourceforge.net.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *Innovations in Theoretical Computer Science Conference*, pages 214–226, 4 2011.

Center on Budget and Policy Priorities. Tracking the covid-19 economy's effects on food, housing, and employment hardships. 2020. URL https://www.cbpp.org/research/poverty-and-inequality/tracking-the-covid-19-economys-effects-on-food-housing-and.

Viani B Djeundje, Jonathan Crook, Raffaella Calabrese, and Mona Hamid. Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163:113766, 2021.

Leonardo Gambacorta, Yiping Huang, Han Qiu, and Jingyi Wang. How do machine learning and non-traditional data affect credit scoring? new evidence from a chinese fintech firm. *Bank for International Settlements*, 2019.

María Óskarsdóttir, Cristián Bravo, Carlos Sarraute, Jan Vanthienen, and Bart Baesens. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing Journal*, 74:26–39, 11 2019.

Donncha Marron. 'lending by numbers': Credit scoring and the constitution of risk within american consumer credit. *Economy and Society*, 36:103–133, 2 2007.

Loretta J Mester. What's the point of credit scoring? *Federal Reserve Bank of Philadelphia: Business Review*, pages 3–16, 1997.

Appendix

Gambling Terms	
bet365	national lottery
bet*	lotto
bingo	paddy power
casino	rainbow riches
gala bingo	skybet
heartgames	sky bet*
jackpotjoy	tombola
lottery	virgingames
mecca	william hill
monopoly	who internet

Figure A.1: Gambling-related terms used for NLP processing of transaction references. *Note: If outgoing transaction references contained any of these terms in its text, they are considered gambling-related.*

Credit Card Providers	
Traditional	*Non-traditional*
Capital One	Vanquis
Barclaycard	Aqua
Hsbc	Marbles
Halifax	Newday
American Express	Fluid
Virgin	Tandem
Tesco	
Mbna	
Lloyds	
Natwest	
Santander	
M&S	
Rbs	
Classic	

Loan Providers		
Traditional	*Non-traditional*	
Capital One	Clearpay	Piggybank
Tesco	Safetynet	Moneyboat
Natwest	Lowell	First Response
Rac	Klarna	Cash4Unow
Barclays	Vanquis	Ferratum
Hsbc	Amigo	Creditspring
Mbna	Lendingstream	Capquest
Santander	Sunny	Lendable
Vauxhall Finance	Brighthouse	Tappily
Nationwide	Laybuy	1Plus1
Rbs	118 118	Loansdirect
Aa	Morses Club	Harvey & Thompson
Lloyds	Provident	Fredrickson
Bmw Finance	Satsuma	Wageday
Virgin	Cabot	Premium Credit
	247 Moneybox	Equita
	Drafty	Wonga
	Loans2Go	Dot Dot
	Very	Ratesetter
	Everyday	Elderbridge
	Likely	Cashplus
	Mrlender	Stepchange
	Creation	Buddy
	Myjar	Rossendales
	Newday	Payplan
	Moneybarn	Zopa
	Moorcroft	Quidie
	Bamboo	Fernovo
	Unclebuck	Moneyline
	Oakam	Cashfloat
	Loansathome	Poundstopocket
	Wescott	Akinika
	Snap	Avant
	Advantage	Quidmarket
	Peachy	Granitefinance
	Zilch	Swiftsterling
	Quick Quid	Novaloans
	Naylors	Bwlegal

Figure A.2: Selection of credit card and loan providers based on traditional (high-street banks) or non-traditional (newer banks). *Note: If incoming (credit) or outgoing (repayment) transaction references contained these terms in its text, they are correspondingly categorized.*

Section	Category	Variable	Stat	Gender	Age Range									
				Female	18 - 21	22 - 25	26 - 30	31 - 35	36 - 40	41 - 45	46 - 50	51 - 55	56 - 60	60+
		No. of Applicants		45,347	2,530	6,127	10,402	11,600	10,293	7,774	5,679	4,144	1,914	855
		%		74.0%	4.1%	10.0%	17.0%	18.9%	16.8%	12.7%	9.3%	6.8%	3.1%	1.4%
	Profile	Female	mean		0.67	0.72	0.71	0.74	0.76	0.76	0.76	0.76	0.75	0.70
			std		0.47	0.45	0.45	0.44	0.43	0.43	0.43	0.42	0.43	0.46
		Age range	mean	3.77										
			std	2.06										
		Employment length (bucket)	mean	2.87	1.86	2.30	2.63	2.90	3.03	3.13	3.21	3.30	3.40	3.37
			std	1.26	1.11	1.25	1.29	1.23	1.20	1.19	1.15	1.10	1.03	1.09
	Residential Status	Council tenant (binary)	mean	0.16	0.05	0.08	0.12	0.15	0.17	0.18	0.18	0.15	0.18	0.17
			std	0.37	0.22	0.27	0.33	0.36	0.38	0.38	0.38	0.36	0.39	0.37
		Housing association tenant (binary)	mean	0.22	0.14	0.17	0.20	0.23	0.23	0.23	0.22	0.23	0.24	0.22
			std	0.41	0.35	0.38	0.40	0.42	0.42	0.42	0.41	0.42	0.43	0.42
		Living with parents (binary)	mean	0.14	0.62	0.44	0.26	0.14	0.09	0.06	0.05	0.04	0.02	0.04
			std	0.35	0.49	0.50	0.44	0.35	0.28	0.24	0.21	0.19	0.15	0.19
		Owner occupier (binary)	mean	0.15	0.02	0.04	0.07	0.11	0.15	0.20	0.25	0.29	0.30	0.32
			std	0.36	0.14	0.19	0.25	0.31	0.36	0.40	0.43	0.46	0.46	0.47
		Private tenant (binary)	mean	0.32	0.17	0.27	0.35	0.37	0.36	0.33	0.30	0.29	0.25	0.26
			std	0.47	0.38	0.44	0.48	0.48	0.48	0.47	0.46	0.46	0.43	0.44
	Inflow	Total income	mean	2944	1975	2453	2860	3085	3166	3165	3101	3056	3025	3039
			std	1521	1060	1301	1488	1552	1598	1617	1588	1634	1628	1767
		Salary-based income	mean	1462	993	1182	1337	1469	1589	1700	1790	1841	1804	1712
			std	910	685	762	826	893	953	980	964	1001	946	977
		Monthly salary consistency (binary)	mean	0.84	0.76	0.78	0.81	0.82	0.84	0.85	0.86	0.86	0.88	0.84
			std	0.37	0.43	0.42	0.39	0.38	0.36	0.36	0.34	0.35	0.33	0.37
Financial Management	Outflow (Static)	Total expenditure	mean	3,626	2,553	3,089	3,546	3,781	3,858	3,835	3,727	3,683	3,608	3,688
			std	1,868	1,352	1,649	1,859	1,919	1,959	1,950	1,892	1,987	1,935	2,194
		Fixed expenditure	mean	2,211	1,542	1,907	2,190	2,292	2,323	2,340	2,260	2,269	2,182	2,277
			std	1,450	1,026	1,273	1,433	1,443	1,489	1,531	1,490	1,569	1,519	1,725
		Flexible expenditure	mean	1,415	1,011	1,182	1,356	1,488	1,535	1,495	1,468	1,414	1,426	1,411
			std	864	586	759	886	948	956	918	911	931	948	1,032
		Average transaction	mean	39.9	31.3	36.0	40.3	41.3	42.0	43.4	44.1	46.3	46.8	49.4
			std	21.1	15.9	18.5	22.3	22.8	22.0	23.4	24.7	26.3	23.6	28.6
	Outflow (Categorical)	Charity & donation	mean	2.1	0.8	0.9	1.2	2.1	1.9	2.0	2.8	2.7	3.3	6.4
			std	19.2	7.2	10.6	13.3	23.1	14.1	12.1	27.6	16.0	30.0	41.0
		Child & school	mean	10.2	0.6	4.1	6.8	13.2	12.9	9.7	6.0	3.8	1.5	2.0
			std	45.0	7.0	39.7	42.3	55.8	46.1	33.8	22.8	24.8	9.6	15.2
		Eating out & takeaways	mean	79.7	107.7	96.9	91.1	85.3	79.5	67.8	55.8	42.9	35.7	30.0
			std	76.1	74.9	74.1	86.2	78.7	77.1	81.9	65.5	59.0	48.7	51.0
		Fashion & beauty	mean	106.9	88.2	94.8	91.4	96.3	97.1	97.4	91.4	82.5	75.2	69.9
			std	136.6	79.8	159.9	114.7	134.5	117.9	142.1	134.9	122.2	96.1	129.6
		Fun & leisure	mean	55.8	41.4	45.2	52.8	58.3	60.3	53.1	50.2	47.4	48.5	40.3
			std	101.6	59.2	130.7	108.8	84.4	92.1	77.9	108.9	87.6	85.4	93.6
		Groceries & housekeeping	mean	418.0	217.4	268.4	331.0	390.1	431.2	437.8	430.6	419.4	409.9	385.5
			std	286.7	155.4	193.9	236.9	275.8	300.9	312.2	297.3	324.3	323.6	332.1
		Health & fitness	mean	1.9	2.0	2.3	2.1	2.1	2.0	2.0	2.0	1.2	0.9	1.1
			std	13.8	5.8	7.4	8.4	13.7	10.7	13.4	25.1	8.0	5.0	7.0
		Housing	mean	108.0	31.8	60.0	82.0	98.6	112.4	121.4	132.9	127.9	129.1	120.7
			std	218.8	107.0	155.0	186.2	210.9	222.7	231.1	251.1	243.2	256.2	256.0
		Medical & health	mean	5.9	2.5	3.9	4.2	4.8	5.6	6.3	7.1	7.7	7.7	6.3
			std	22.5	7.0	11.1	12.4	14.8	21.9	18.6	29.5	39.9	19.4	12.6
		Subscriptions	mean	5.0	1.4	2.7	3.7	4.8	5.1	5.5	6.2	5.8	4.9	4.6
			std	14.0	4.6	12.9	9.2	13.2	13.2	14.3	18.7	16.5	11.3	12.9
		Transport & fuel	mean	73.7	79.8	81.5	79.9	78.3	74.8	75.2	68.6	63.8	58.7	52.2
			std	85.4	75.7	79.6	88.7	87.8	82.6	97.8	92.2	82.8	69.4	78.4
		Utilities total	mean	209.8	110.0	144.5	171.8	203.3	222.9	235.8	233.0	222.6	217.9	204.2
			std	160.5	107.4	119.3	130.4	155.5	174.1	178.3	176.5	175.5	162.6	162.1
	Outflow (Gambling)	Gambling expenditure	mean	144.3	90.7	151.2	208.3	225.1	199.7	158.8	168.9	156.4	172.2	173.7
			std	498.3	358.5	478.2	598.2	660.0	615.0	531.8	577.7	545.0	598.8	529.1
		Gambling income	mean	30.4	25.6	38.7	50.1	52.6	44.5	33.0	34.8	29.1	36.8	31.7
			std	167.9	137.4	180.0	219.5	246.9	216.7	179.5	192.7	167.2	223.8	196.9
		No. of gambling transactions	mean	9	5	8	10	12	10	9	9	8	9	9
			std	28	16	26	28	32	29	28	30	27	28	25
	Outflow (Temporal)	Daily persistence in fixed expenditure	mean	0.37	0.36	0.36	0.36	0.37	0.37	0.36	0.35	0.35	0.34	0.33
			std	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.13	0.14	0.14	0.13
		Daily persistence in flexible expenditure	mean	0.46	0.43	0.43	0.44	0.45	0.46	0.46	0.45	0.44	0.44	0.43
			std	0.15	0.13	0.14	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.15
		Monthly persistence in categorical spending	mean	0.76	0.70	0.72	0.74	0.75	0.77	0.77	0.78	0.78	0.77	0.78
			std	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.11	0.11
		Burstiness in fixed expenditure	mean	0.27	0.27	0.28	0.28	0.28	0.28	0.27	0.27	0.27	0.27	0.26
			std	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11
		Burstiness in flexible expenditure	mean	0.35	0.38	0.38	0.37	0.36	0.36	0.34	0.33	0.32	0.31	0.30
			std	0.11	0.10	0.11	0.11	0.12	0.11	0.11	0.12	0.12	0.12	0.12

Section	Category	Variable	Stat	Benefits Received		
				Carer	Child	Disability
		No. of Applicants		1,524	22,973	4,875
		%		2.5%	37.5%	8.0%
	Profile	Female	mean	0.85	0.95	0.86
			std	0.36	0.23	0.35
		Age range	mean	3.71	3.69	4.23
			std	1.86	1.56	1.87
		Employment length (bucket)	mean	2.64	2.92	2.92
			std	1.36	1.25	1.26
	Residential Status	Council tenant (binary)	mean	0.24	0.21	0.23
			std	0.42	0.40	0.42
		Housing association tenant (binary)	mean	0.32	0.25	0.28
			std	0.47	0.43	0.45
		Living with parents (binary)	mean	0.10	0.06	0.06
			std	0.30	0.24	0.24
		Owner occupier (binary)	mean	0.09	0.15	0.14
			std	0.29	0.36	0.35
		Private tenant (binary)	mean	0.26	0.33	0.29
			std	0.44	0.47	0.45
	Inflow	Total income	mean	3223	3150	3453
			std	1779	1555	1683
		Salary-based income	mean	783	1389	1235
			std	880	931	961
		Monthly salary consistency (binary)	mean	0.57	0.83	0.76
			std	0.50	0.38	0.43
Financial Management	Outflow (Static)	Total expenditure	mean	4,202	3,918	4,313
			std	2,289	1,914	2,094
		Fixed expenditure	mean	2,419	2,341	2,544
			std	1,792	1,499	1,645
		Flexible expenditure	mean	1,784	1,577	1,769
			std	970	874	954
		Average transaction	mean	37.1	38.1	40.2
			std	19.6	18.6	22.3
	Outflow (Categorical)	Charity & donation	mean	2.2	2.1	2.5
			std	18.6	19.9	16.8
		Child & school	mean	11.5	16.9	12.1
			std	49.9	58.5	41.9
		Eating out & takeaways	mean	106.2	90.5	99.0
			std	100.9	82.6	94.0
		Fashion & beauty	mean	105.5	114.8	116.4
			std	123.9	140.5	153.2
		Fun & leisure	mean	65.1	62.6	70.4
			std	85.4	92.1	97.3
		Groceries & housekeeping	mean	524.5	493.6	540.3
			std	370.4	302.6	352.0
		Health & fitness	mean	1.2	1.8	1.9
			std	5.2	10.8	10.0
		Housing	mean	85.8	121.0	105.9
			std	186.8	226.8	207.7
		Medical & health	mean	4.7	5.5	6.5
			std	10.8	16.4	17.9
		Subscriptions	mean	2.6	5.4	4.7
			std	7.6	15.6	12.8
		Transport & fuel	mean	77.1	78.9	83.8
			std	100.9	87.6	90.2
		Utilities total	mean	233.1	245.4	260.0
			std	192.9	171.7	185.7
	Outflow (Gambling)	Gambling expenditure	mean	213.3	150.8	186.2
			std	564.1	500.2	562.6
		Gambling income	mean	48.0	32.5	39.9
			std	199.5	169.3	189.1
		No. of gambling transactions	mean	12	10	11
			std	30	29	30
	Outflow (Temporal)	Daily persistence in fixed expenditure	mean	0.43	0.40	0.40
			std	0.13	0.12	0.12
		Daily persistence in flexible expenditure	mean	0.49	0.49	0.48
			std	0.14	0.14	0.14
		Monthly persistence in categorical spending	mean	0.78	0.78	0.78
			std	0.11	0.10	0.10
		Burstiness in fixed expenditure	mean	0.28	0.27	0.28
			std	0.10	0.09	0.10
		Burstiness in flexible expenditure	mean	0.37	0.35	0.36
			std	0.11	0.11	0.11

Section	Category	Variable	Stat	Financial Vulnerability Indicators*									
				Financial shock withstanding	Financial shock withstanding *never*	Financial shock withstanding *always*	Gambler	Insolvent	Insufficient disposable income	Overdraft	Overdraft *always*	Overdraft *never*	Returned Direct Debits
		No. of Applicants		29,639	11,290	11,666	12,394	2,460	7,128	41,328	21,530	4,215	17,371
		%		48.3%	18.4%	19.0%	20.2%	4.0%	11.6%	67.4%	35.1%	6.9%	28.3%
	Profile	Female	mean	0.73	0.73	0.78	0.60	0.75	0.72	0.74	0.74	0.75	0.76
			std	0.45	0.44	0.41	0.49	0.44	0.45	0.44	0.44	0.43	0.43
		Age range	mean	3.38	3.36	4.45	3.59	4.21	3.80	3.62	3.61	4.39	3.58
			std	2.06	2.03	1.96	1.95	1.92	2.16	2.03	2.02	2.12	1.90
		Employment length (bucket)	mean	2.77	2.83	3.12	2.93	2.98	2.86	2.86	2.91	3.06	2.78
			std	1.29	1.28	1.16	1.24	1.23	1.27	1.27	1.25	1.19	1.29
	Residential Status	Council tenant (binary)	mean	0.13	0.12	0.16	0.15	0.16	0.14	0.14	0.13	0.18	0.16
			std	0.34	0.33	0.37	0.35	0.37	0.35	0.35	0.34	0.38	0.36
		Housing association tenant (binary)	mean	0.21	0.21	0.21	0.22	0.23	0.20	0.21	0.21	0.21	0.24
			std	0.41	0.41	0.41	0.41	0.42	0.40	0.41	0.41	0.41	0.43
		Living with parents (binary)	mean	0.20	0.20	0.10	0.21	0.10	0.18	0.17	0.18	0.13	0.14
			std	0.40	0.40	0.31	0.41	0.30	0.38	0.38	0.38	0.34	0.35
		Owner occupier (binary)	mean	0.14	0.15	0.18	0.13	0.14	0.14	0.15	0.16	0.15	0.14
			std	0.34	0.36	0.39	0.34	0.35	0.34	0.36	0.37	0.35	0.35
		Private tenant (binary)	mean	0.32	0.32	0.34	0.29	0.36	0.34	0.33	0.32	0.33	0.32
			std	0.47	0.46	0.47	0.45	0.48	0.47	0.47	0.47	0.47	0.47
	Inflow	Total income	mean	2794	2761	3348	3552	2938	2551	3028	3025	2703	3170
			std	1488	1404	1687	1883	1491	1250	1565	1543	1487	1645
		Salary-based income	mean	1379	1390	1846	1501	1654	1408	1513	1541	1643	1467
			std	910	912	986	971	892	807	934	933	922	967
		Monthly salary consistency (binary)	mean	0.80	0.79	0.88	0.80	0.87	0.84	0.83	0.83	0.85	0.81
			std	0.40	0.41	0.33	0.40	0.33	0.37	0.38	0.37	0.35	0.39
Financial Management	Outflow (Static)	Total expenditure	mean	3,397	3,309	4,119	4,422	3,732	2,882	3,659	3,648	3,491	3,893
			std	1,817	1,738	2,031	2,255	1,862	1,393	1,920	1,897	1,846	2,054
		Fixed expenditure	mean	2,170	2,121	2,314	2,342	2,329	1,847	2,264	2,285	2,059	2,452
			std	1,456	1,387	1,495	1,573	1,460	1,093	1,475	1,472	1,457	1,595
		Flexible expenditure	mean	1,228	1,187	1,805	2,080	1,403	1,036	1,395	1,363	1,432	1,441
			std	792	759	1,052	1,277	795	592	901	879	886	954
		Average transaction	mean	40.8	37.6	40.8	39.7	42.5	40.6	41.7	40.7	41.8	43.0
			std	23.7	21.2	20.4	22.8	20.4	21.5	22.6	22.8	25.8	23.0
	Outflow (Categorical)	Charity & donation	mean	1.7	1.4	2.7	1.6	2.1	1.3	2.0	2.0	2.3	2.3
			std	16.6	11.2	20.4	20.3	27.2	4.1	18.1	18.6	23.0	22.5
		Child & school	mean	7.0	6.9	11.7	6.6	11.1	6.7	8.8	8.9	7.5	10.1
			std	36.8	33.0	47.7	38.3	40.1	30.7	42.4	44.0	34.3	48.0
		Eating out & takeaways	mean	72.0	70.8	90.6	72.4	75.1	61.2	76.5	73.8	74.6	77.9
			std	75.3	71.2	84.8	77.4	69.6	62.2	76.5	74.5	95.6	78.5
		Fashion & beauty	mean	82.7	83.7	116.8	79.5	89.8	69.7	92.6	92.3	92.5	91.2
			std	127.6	131.5	123.3	135.9	115.9	90.0	131.9	128.5	109.0	137.2
		Fun & leisure	mean	44.4	43.0	70.5	52.1	58.3	37.1	52.3	50.5	55.2	54.9
			std	91.0	73.5	105.3	106.1	74.9	53.6	94.5	101.1	75.6	112.8
		Groceries & housekeeping	mean	319.9	318.1	528.7	324.3	427.5	330.2	370.0	362.5	428.8	373.6
			std	247.6	246.5	332.6	276.6	290.6	245.3	272.5	265.0	318.4	279.1
		Health & fitness	mean	1.8	1.8	2.2	1.8	2.1	1.7	2.1	2.1	1.5	2.5
			std	9.3	8.0	20.3	11.3	9.5	9.0	9.3	9.0	9.3	11.0
		Housing	mean	87.9	86.1	127.7	82.4	131.8	100.7	102.0	103.5	107.4	106.7
			std	204.3	201.6	235.9	195.1	236.3	202.9	217.3	222.2	218.1	220.8
		Medical & health	mean	4.5	4.6	7.8	3.9	6.2	4.4	5.2	5.2	6.5	5.0
			std	19.2	26.6	21.6	15.5	23.6	11.2	18.0	15.6	20.6	13.8
		Subscriptions	mean	4.2	4.5	6.2	3.4	6.0	4.1	4.8	5.1	4.1	5.0
			std	11.7	12.4	18.5	11.2	27.4	11.0	12.5	12.4	9.9	12.6
		Transport & fuel	mean	70.5	71.4	85.9	65.8	76.0	64.3	79.7	81.4	62.6	82.2
			std	85.3	87.9	91.8	80.0	81.0	69.8	88.5	90.3	80.8	91.1
		Utilities total	mean	180.9	183.7	248.1	184.6	236.9	182.0	209.5	213.0	185.8	251.4
			std	154.8	158.3	165.8	165.4	158.5	143.2	165.5	170.9	156.0	184.9
	Outflow (Gambling)	Gambling expenditure	mean	129.2	103.4	254.3	863.3	150.1	55.6	180.6	173.5	149.4	206.4
			std	444.4	373.7	755.0	1,044.9	471.3	210.6	575.5	570.1	490.0	625.5
		Gambling income	mean	29.8	24.4	57.8	196.5	32.9	12.3	40.8	38.9	33.7	47.8
			std	159.9	139.8	272.7	427.7	189.8	76.3	206.2	203.4	177.4	222.4
		No. of gambling transactions	mean	7	7	13	43	8	4	9	9	9	11
			std	22	22	41	51	25	14	27	27	28	30
	Outflow (Temporal)	Daily persistence in fixed expenditure	mean	0.37	0.39	0.37	0.36	0.35	0.34	0.36	0.38	0.36	0.36
			std	0.13	0.13	0.13	0.14	0.13	0.13	0.13	0.13	0.14	0.13
		Daily persistence in flexible expenditure	mean	0.43	0.43	0.51	0.48	0.46	0.41	0.43	0.43	0.50	0.41
			std	0.15	0.16	0.13	0.15	0.15	0.16	0.15	0.16	0.15	0.15
		Monthly persistence in categorical spending	mean	0.74	0.75	0.80	0.77	0.77	0.75	0.75	0.75	0.78	0.74
			std	0.12	0.12	0.09	0.11	0.10	0.12	0.11	0.11	0.11	0.11
		Burstiness in fixed expenditure	mean	0.28	0.27	0.26	0.29	0.27	0.27	0.28	0.28	0.26	0.28
			std	0.09	0.09	0.10	0.10	0.09	0.10	0.10	0.09	0.10	0.10
		Burstiness in flexible expenditure	mean	0.36	0.35	0.34	0.44	0.34	0.34	0.36	0.37	0.32	0.37
			std	0.12	0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.12	0.11

Figure A.3: Summary statistics of applicants' pool (Part 1). *Note: Table reports the mean and SD (in parentheses) of the variables. Profile-related features are provided as inputs by the applicants themselves during the application process. All other features are constructed based on the transaction data collected at the time of application via a third-party Open Banking API provider. All attributes represent monthly averages with amount values in £/GBP unless noted otherwise.*

			Gender	Age Range							Benefits Received			Financial Vulnerability Indicators *									
			Female	18 - 21	22 - 25	26 - 30	31 - 35	36 - 40	41 - 45	46 - 50	51 - 55	56 - 60	60+	Carer	Child	Disability	Financial shock withstanding	Financial shock withstanding *never*	Financial shock withstanding *always*	Gambler	Insolvent	Insufficient disposable income	Overdraft	Overdraft *always*	Overdraft *never*	Returned Direct Debits
Financial Management	Volatility	Volatility: Account balance	mean	0.44	0.55	0.49	0.46	0.45	0.43	0.43	0.43	0.46	0.46	0.44	0.43	0.43	0.51	0.38	0.34	0.49	0.38	0.44	0.49	0.49	0.35	0.45
			std	*0.23*	*0.24*	*0.24*	*0.23*	*0.22*	*0.22*	*0.23*	*0.23*	*0.24*	*0.24*	*0.21*	*0.22*	*0.21*	*0.27*	*0.26*	*0.14*	*0.23*	*0.18*	*0.23*	*0.25*	*0.28*	*0.15*	*0.22*
		Volatility: Total expenditure	mean	0.33	0.40	0.37	0.35	0.33	0.33	0.32	0.32	0.32	0.34	0.32	0.32	0.31	0.34	0.32	0.31	0.38	0.31	0.32	0.34	0.33	0.32	0.32	0.33
			std	*0.14*	*0.14*	*0.14*	*0.14*	*0.14*	*0.14*	*0.14*	*0.14*	*0.16*	*0.15*	*0.14*	*0.13*	*0.14*	*0.14*	*0.13*	*0.14*	*0.15*	*0.13*	*0.13*	*0.14*	*0.14*	*0.14*	*0.14*
		Volatility: Fixed expenditure	mean	0.38	0.47	0.43	0.40	0.39	0.38	0.37	0.37	0.37	0.37	0.39	0.37	0.37	0.39	0.37	0.37	0.42	0.35	0.37	0.39	0.38	0.38	0.38	0.38
			std	*0.15*	*0.15*	*0.15*	*0.15*	*0.15*	*0.16*	*0.16*	*0.16*	*0.17*	*0.17*	*0.16*	*0.15*	*0.15*	*0.16*	*0.15*	*0.16*	*0.16*	*0.15*	*0.15*	*0.15*	*0.15*	*0.17*	*0.15*
		Volatility: Flexible expenditure	mean	0.41	0.48	0.46	0.45	0.43	0.41	0.41	0.41	0.41	0.43	0.40	0.39	0.39	0.45	0.43	0.37	0.47	0.40	0.43	0.43	0.43	0.43	0.40	0.44
			std	*0.14*	*0.15*	*0.16*	*0.16*	*0.16*	*0.16*	*0.15*	*0.15*	*0.16*	*0.16*	*0.16*	*0.15*	*0.15*	*0.16*	*0.16*	*0.14*	*0.16*	*0.15*	*0.15*	*0.15*	*0.15*	*0.16*	*0.15*
		Volatility: Income	mean	0.32	0.40	0.37	0.35	0.33	0.32	0.31	0.31	0.31	0.34	0.32	0.32	0.31	0.34	0.31	0.31	0.39	0.30	0.31	0.33	0.32	0.32	0.32	0.33
			std	*0.16*	*0.17*	*0.16*	*0.16*	*0.16*	*0.16*	*0.16*	*0.16*	*0.18*	*0.17*	*0.15*	*0.15*	*0.15*	*0.16*	*0.15*	*0.16*	*0.16*	*0.16*	*0.15*	*0.16*	*0.15*	*0.17*	*0.15*
		Volatility: Salary	mean	0.32	0.43	0.39	0.36	0.33	0.31	0.29	0.27	0.26	0.27	0.50	0.34	0.38	0.34	0.32	0.28	0.35	0.30	0.30	0.33	0.31	0.29	0.29	0.35
			std	*0.26*	*0.30*	*0.29*	*0.27*	*0.26*	*0.25*	*0.25*	*0.23*	*0.23*	*0.24*	*0.34*	*0.27*	*0.30*	*0.28*	*0.28*	*0.22*	*0.27*	*0.24*	*0.26*	*0.26*	*0.26*	*0.34*	*0.27*
Financial Distress		Buy Now Pay Later financing (BNPL)	mean	50.7	35.8	44.1	45.1	48.5	48.4	40.8	37.0	29.6	23.8	23.0	62.5	58.4	62.0	37.7	39.7	55.6	35.7	41.2	34.0	42.3	43.5	43.6	42.8
			std	*90.5*	*56.8*	*79.3*	*79.2*	*88.4*	*94.5*	*83.1*	*80.1*	*68.8*	*62.2*	*75.7*	*116.9*	*98.8*	*109.9*	*74.9*	*80.5*	*101.0*	*78.1*	*79.8*	*67.3*	*81.1*	*83.5*	*74.2*	*92.8*
		Insolvency & debt management (DM)	mean	28.1	6.6	17.3	25.3	28.1	29.9	32.1	33.4	36.3	34.9	34.7	21.3	29.6	27.1	22.2	21.0	39.2	27.2	103.7	24.8	27.1	25.8	33.2	31.3
			std	*59.7*	*24.8*	*38.5*	*51.3*	*56.9*	*59.3*	*72.9*	*72.6*	*79.7*	*76.7*	*68.9*	*50.9*	*59.8*	*52.5*	*52.2*	*45.6*	*75.9*	*65.8*	*100.2*	*56.4*	*58.1*	*51.5*	*74.2*	*63.7*
		No. of days in Overdraft (OD)	mean	6.2	6.0	7.6	6.8	6.3	5.8	6.0	5.9	5.7	5.6	5.7	5.0	6.3	5.6	10.9	15.9	0.8	6.1	3.0	6.4	9.1	14.1	0.0	7.4
			std	*7.5*	*7.1*	*8.0*	*7.7*	*7.4*	*7.5*	*7.4*	*7.3*	*7.3*	*7.1*	*6.8*	*7.6*	*7.2*	*8.2*	*7.7*	*7.7*	*1.5*	*7.5*	*4.9*	*7.5*	*6.8*	*0.0*	*7.4*
		No. of Returned Direct Debits (RDD)	mean	0.87	0.57	0.77	0.92	0.92	0.88	0.87	0.79	0.67	0.60	0.54	1.10	0.99	1.05	1.03	1.01	0.36	0.87	0.87	0.81	1.13	1.19	0.02	2.40
			std	*1.42*	*0.89*	*1.10*	*1.35*	*1.48*	*1.41*	*1.43*	*1.45*	*1.30*	*1.16*	*1.04*	*1.66*	*1.51*	*1.65*	*1.52*	*1.49*	*0.78*	*1.41*	*1.38*	*1.29*	*1.54*	*1.75*	*0.12*	*1.70*
Financial Resilience		Account Balance: Mean	mean	196	99	59	121	186	233	229	278	298	342	333	211	204	255	-139	-521	676	229	384	148	50	-206	626	120
			std	*536*	*382*	*486*	*541*	*527*	*556*	*595*	*597*	*608*	*657*	*638*	*445*	*520*	*508*	*559*	*632*	*378*	*561*	*358*	*546*	*574*	*636*	*452*	*500*
		Account Balance: Minimum	mean	-264	-219	-315	-302	-270	-260	-284	-263	-284	-245	-237	-157	-250	-206	-539	-905	39	-268	-83	-282	-425	-491	98	-320
			std	*535*	*384*	*508*	*554*	*528*	*569*	*600*	*577*	*588*	*615*	*593*	*417*	*511*	*481*	*681*	*805*	*205*	*559*	*273*	*541*	*613*	*709*	*234*	*529*
		Account Balance: Maximum	mean	1,511	1,080	1,202	1,407	1,540	1,640	1,677	1,745	1,771	1,817	1,780	1,317	1,515	1,555	1,151	758	2,160	1,661	1,776	1,345	1,417	1,183	1,957	1,448
			std	*887*	*605*	*757*	*840*	*861*	*954*	*973*	*1,003*	*1,061*	*1,091*	*1,257*	*818*	*890*	*894*	*822*	*785*	*913*	*975*	*789*	*930*	*907*	*925*	*856*
		No. of transactions	mean	140	129	138	141	145	140	132	125	114	109	107	182	157	166	137	142	138	179	133	106	137	138	121	150
			std	*79*	*65*	*78*	*79*	*83*	*78*	*73*	*71*	*64*	*61*	*62*	*95*	*83*	*89*	*76*	*77*	*80*	*102*	*76*	*54*	*76*	*75*	*72*	*86*
		% of M can withstand financial shock	mean	0.52	0.37	0.38	0.45	0.51	0.56	0.56	0.59	0.61	0.62	0.61	0.53	0.54	0.59	0.16	0.00	1.00	0.57	0.68	0.48	0.39	0.21	0.87	0.41
			std	*0.38*	*0.33*	*0.35*	*0.37*	*0.38*	*0.38*	*0.38*	*0.38*	*0.38*	*0.38*	*0.38*	*0.39*	*0.39*	*0.39*	*0.16*	*0.00*	*0.00*	*0.49*	*0.34*	*0.38*	*0.35*	*0.28*	*0.25*	*0.35*
		Deposable income	mean	771	597	704	781	822	819	781	760	736	761	785	1,018	836	931	731	704	884	1,165	768	50	756	742	799	821
			std	*966*	*649*	*806*	*890*	*946*	*952*	*902*	*877*	*902*	*919*	*985*	*1,130*	*927*	*1,030*	*853*	*809*	*1,006*	*1,140*	*876*	*29*	*866*	*866*	*916*	*966*
Financial Planning		Has any insurance and/or pension (binary)	mean	0.75	0.55	0.61	0.68	0.73	0.75	0.77	0.80	0.82	0.83	0.82	0.73	0.79	0.79	0.68	0.67	0.82	0.67	0.82	0.71	0.76	0.76	0.68	0.82
			std	*0.43*	*0.50*	*0.49*	*0.47*	*0.44*	*0.43*	*0.42*	*0.40*	*0.39*	*0.37*	*0.39*	*0.45*	*0.41*	*0.40*	*0.47*	*0.47*	*0.38*	*0.47*	*0.38*	*0.45*	*0.43*	*0.47*	*0.43*	*0.38*
		Insurance & pension spending	mean	47.8	41.1	44.4	44.2	46.5	48.1	50.2	50.9	55.4	56.7	58.9	44.8	52.2	54.2	44.5	43.9	55.4	41.3	56.0	44.5	53.4	55.1	37.6	67.1
			std	*70.5*	*73.2*	*81.5*	*67.5*	*71.7*	*83.2*	*66.5*	*64.6*	*70.9*	*65.0*	*71.2*	*64.1*	*76.6*	*84.8*	*75.9*	*74.3*	*66.0*	*80.5*	*71.6*	*64.3*	*78.6*	*83.1*	*55.4*	*95.3*
		Has appliance insurance (binary)	mean	0.11	0.02	0.04	0.07	0.10	0.11	0.13	0.13	0.13	0.16	0.14	0.19	0.14	0.18	0.09	0.08	0.13	0.08	0.13	0.09	0.10	0.10	0.10	0.12
			std	*0.32*	*0.15*	*0.20*	*0.26*	*0.30*	*0.32*	*0.33*	*0.34*	*0.34*	*0.37*	*0.35*	*0.39*	*0.35*	*0.39*	*0.28*	*0.28*	*0.34*	*0.27*	*0.33*	*0.28*	*0.30*	*0.31*	*0.30*	*0.33*
		Has life insurance (binary)	mean	0.11	0.05	0.06	0.08	0.10	0.11	0.12	0.13	0.14	0.15	0.17	0.12	0.13	0.12	0.09	0.09	0.13	0.08	0.15	0.10	0.11	0.10	0.10	0.11
			std	*0.31*	*0.23*	*0.24*	*0.27*	*0.30*	*0.31*	*0.33*	*0.34*	*0.35*	*0.36*	*0.37*	*0.32*	*0.33*	*0.32*	*0.29*	*0.28*	*0.34*	*0.28*	*0.36*	*0.30*	*0.31*	*0.31*	*0.30*	*0.31*
		Has pet insurance (binary)	mean	0.17	0.09	0.11	0.13	0.16	0.17	0.17	0.18	0.18	0.15	0.16	0.20	0.19	0.22	0.13	0.12	0.20	0.12	0.20	0.14	0.15	0.15	0.16	0.18
			std	*0.38*	*0.29*	*0.31*	*0.34*	*0.36*	*0.38*	*0.37*	*0.38*	*0.38*	*0.36*	*0.36*	*0.40*	*0.39*	*0.41*	*0.34*	*0.33*	*0.40*	*0.33*	*0.40*	*0.35*	*0.36*	*0.35*	*0.37*	*0.38*
		Savings (% of total income)	mean	331 (.12)	299 (.16)	352 (.15)	398 (.13)	393 (.12)	318 (.10)	297 (.10)	256 (.08)	229 (.08)	245 (.08)	221 (.07)	(.0 437 (.14)	382 (.12)	388 (.11)	424 (.17)	450 (.18)	141 (.04)	379 (.11)	234 (.08)	198 (.08)	375 (.13)	374 (.13)	90 (.03)	510 (.18)
			std	*781 (.28)*	*632 (.36)*	*772 (.34)*	*798 (.30)*	*814 (.28)*	*776 (.26)*	*762 (.26)*	*711 (.23)*	*639 (.22)*	*651 (.24)*	*609 (.23)*	*(.0 957 (.30)*	*836 (.28)*	*908 (.27)*	*856 (.34)*	*847 (.35)*	*449 (.12)*	*892 (.27)*	*649 (.25)*	*551 (.24)*	*828 (.30)*	*828 (.29)*	*347 (.15)*	*973 (.35)*
Financial Aid		Total benefits (% of total income)	mean	411 (.17)	123 (.10)	218 (.12)	356 (.16)	447 (.18)	452 (.17)	359 (.14)	237 (.09)	168 (.07)	123 (.06)	138 (.06)	1388 (.51)	700 (.27)	1144 (.40)	311 (.14)	308 (.14)	385 (.14)	321 (.12)	382 (.15)	265 (.12)	335 (.14)	321 (.13)	299 (.14)	424 (.16)
			std	*559 (.23)*	*274 (.20)*	*415 (.22)*	*538 (.23)*	*588 (.23)*	*582 (.22)*	*530 (.21)*	*433 (.17)*	*379 (.17)*	*311 (.15)*	*357 (.17)*	*775 (.27)*	*539 (.23)*	*703 (.25)*	*491 (.22)*	*484 (.22)*	*585 (.21)*	*526 (.20)*	*550 (.22)*	*456 (.20)*	*511 (.20)*	*505 (.23)*	*578 (.22)*
		Carer benefits	mean	6.3	3.7	2.9	5.5	7.0	7.2	5.7	4.3	4.5	4.5	5.8	222.9	9.9	46.7	5.4	5.4	6.9	6.9	6.1	3.8	5.0	4.5	5.9	6.8
			std	*41.0*	*32.1*	*28.2*	*38.4*	*43.3*	*43.7*	*40.9*	*34.9*	*35.7*	*35.9*	*37.6*	*112.8*	*51.1*	*104.4*	*38.6*	*37.8*	*42.7*	*43.8*	*42.9*	*31.8*	*37.2*	*34.2*	*39.3*	*43.6*
		Child benefits	mean	93.2	9.5	26.3	60.9	105.0	119.2	95.9	57.6	27.3	10.8	13.8	232.1	195.1	180.0	64.7	68.9	96.0	64.4	87.4	57.1	74.8	75.3	69.1	90.7
			std	*171.3*	*37.2*	*73.1*	*130.7*	*185.5*	*204.3*	*179.4*	*132.3*	*88.6*	*55.8*	*100.4*	*341.0*	*206.1*	*296.5*	*142.9*	*153.0*	*191.0*	*148.6*	*156.3*	*137.4*	*160.3*	*162.0*	*158.1*	*178.2*
		Disability benefits	mean	34.1	7.9	10.0	21.5	29.0	38.9	40.4	35.6	34.0	36.8	39.4	275.2	46.5	389.9	23.7	22.9	47.0	30.8	36.0	23.8	28.6	27.1	29.9	36.0
			std	*123.5*	*59.6*	*65.5*	*96.2*	*111.0*	*130.6*	*134.4*	*130.1*	*132.2*	*130.9*	*124.1*	*268.1*	*141.7*	*201.2*	*101.8*	*98.2*	*147.7*	*119.8*	*121.5*	*102.0*	*112.6*	*109.4*	*119.2*	*127.2*
		Pension income	mean	5.9	0.5	0.2	0.9	0.8	1.4	4.2	7.2	16.7	49.5	155.1	9.8	2.4	12.4	5.1	5.1	11.5	8.7	9.7	6.2	6.3	6.0	11.1	5.6
			std	*68.8*	*19.6*	*5.3*	*25.2*	*22.3*	*30.1*	*55.6*	*80.3*	*120.1*	*190.5*	*317.2*	*89.3*	*44.0*	*101.0*	*62.2*	*60.2*	*98.3*	*87.5*	*91.0*	*69.9*	*70.0*	*66.5*	*97.7*	*66.5*
Financial Inclusion		No. of unique traditional CCs	mean	0.17	0.14	0.16	0.17	0.17	0.19	0.20	0.21	0.23	0.22	0.25	0.11	0.16	0.17	0.19	0.21	0.20	0.20	0.10	0.17	0.20	0.23	0.16	0.18
			std	*0.50*	*0.45*	*0.47*	*0.50*	*0.50*	*0.52*	*0.54*	*0.56*	*0.57*	*0.56*	*0.60*	*0.42*	*0.46*	*0.52*	*0.52*	*0.54*	*0.55*	*0.54*	*0.38*	*0.48*	*0.54*	*0.57*	*0.51*	*0.51*
		No. of unique non-traditional CCs	mean	0.18	0.12	0.14	0.17	0.18	0.18	0.20	0.21	0.23	0.26	0.25	0.12	0.17	0.19	0.18	0.21	0.23	0.18	0.10	0.18	0.19	0.22	0.19	0.17
			std	*0.53*	*0.41*	*0.47*	*0.53*	*0.54*	*0.55*	*0.55*	*0.55*	*0.58*	*0.61*	*0.62*	*0.43*	*0.52*	*0.54*	*0.53*	*0.57*	*0.60*	*0.53*	*0.37*	*0.52*	*0.55*	*0.58*	*0.55*	*0.51*
		Loans received	mean	33.1	16.1	22.3	30.7	34.3	38.7	41.5	46.5	46.8	52.7	59.5	19.5	29.6	34.2	34.2	35.3	46.1	40.0	19.4	30.2	38.4	42.8	40.4	28.9
			std	*101.8*	*90.1*	*75.3*	*100.0*	*107.0*	*121.8*	*116.4*	*130.4*	*123.6*	*121.4*	*148.4*	*76.4*	*96.1*	*101.2*	*107.4*	*108.1*	*132.7*	*125.5*	*66.1*	*87.4*	*113.5*	*119.6*	*120.0*	*94.9*
		No. of unique traditional loans	mean	0.68	0.54	0.57	0.62	0.66	0.67	0.75	0.79	0.81	0.86	0.83	0.51	0.65	0.70	0.65	0.68	0.79	0.66	0.42	0.66	0.72	0.77	0.64	0.67
			std	*1.01*	*0.93*	*0.93*	*0.99*	*1.00*	*1.02*	*1.04*	*1.05*	*1.07*	*1.07*	*1.03*	*0.90*	*1.00*	*1.05*	*0.99*	*1.01*	*1.06*	*1.02*	*0.83*	*0.98*	*1.03*	*1.06*	*1.04*	*0.99*
		No. of unique non-traditional loans	mean	2.61	2.35	2.64	2.51	2.49	2.38	2.28	2.19	2.02	1.80	1.62	2.19	2.69	2.52	2.24	2.35	2.71	1.94	1.98	2.27	2.38	2.63	2.48	2.07
			std	*2.96*	*2.67*	*2.85*	*2.79*	*2.93*	*2.85*	*2.82*	*2.75*	*2.57*	*2.30*	*2.31*	*2.57*	*3.02*	*2.96*	*2.68*	*2.73*	*3.17*	*2.54*	*2.58*	*2.64*	*2.73*	*2.98*	*3.28*	*2.18*
		Loans received	mean	182	150	201	194	190	188	189	199	202	194	199	108	168	167	190	196	204	190	119	170	205	224	155	181
			std	*303*	*241*	*310*	*310*	*315*	*312*	*315*	*317*	*314*	*322*	*381*	*236*	*293*	*293*	*305*	*306*	*333*	*331*	*258*	*240*	*322*	*329*	*281*	*297*
		Loan payments	mean	315	188	270	300	317	325	343	366	382	364	376	232	308	327	307	332	378	302	232	291	343	381	275	330
			std	*368*	*258*	*321*	*357*	*370*	*380*	*395*	*420*	*434*	*403*	*495*	*314*	*355*	*379*	*379*	*399*	*413*	*386*	*308*	*321*	*382*	*415*	*343*	*378*
		Pay/day loan usage (binary)	mean	0.18	0.18	0.22	0.19	0.19	0.17	0.16	0.18	0.18	0.17	0.16	0.10	0.16	0.15	0.21	0.23	0.14	0.14	0.12	0.21	0.20	0.23	0.11	0.22
			std	*0.39*	*0.38*	*0.42*	*0.40*	*0.39*	*0.38*	*0.37*	*0.38*	*0.38*	*0.37*	*0.37*	*0.29*	*0.37*	*0.36*	*0.41*	*0.42*	*0.34*	*0.34*	*0.33*	*0.40*	*0.40*	*0.42*	*0.32*	*0.41*
		Pay/day loan payments	mean	50.6	35.2	51.5	50.9	52.2	49.8	52.8	61.4	64.0	60.1	69.5	25.9	45.6	48.0	58.1	70.2	48.5	39.9	35.4	48.3	59.3	72.0	33.6	55.4
			std	*179.3*	*119.0*	*156.5*	*169.8*	*183.1*	*175.6*	*188.0*	*210.9*	*220.0*	*218.3*	*271.0*	*139.6*	*170.2*	*187.6*	*184.6*	*206.3*	*202.4*	*171.5*	*158.7*	*147.6*	*192.2*	*210.9*	*165.5*	*177.2*

Figure A.4: Summary statistics of applicants' pool (Part 2). *Note: Table reports the mean and SD (in parentheses) of the variables. Profile-related features are provided as inputs by the applicants themselves during the application process. All other features are constructed based on the transaction data collected at the time of application via a third-party Open Banking API provider. All attributes represent monthly averages with amount values in £/GBP unless noted otherwise.*