Edinburgh Research Explorer

The role of web browsing in credit risk prediction

Citation for published version:
Garzon Rozo, BJ, Crook, J & Andreeva, G 2023, 'The role of web browsing in credit risk prediction', *Decision Support Systems*, vol. 164, 113879. https://doi.org/10.1016/j.dss.2022.113879

Digital Object Identifier (DOI):
10.1016/j.dss.2022.113879

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Decision Support Systems

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 25. Nov. 2022

Decision Support Systems 164 (2023) 113879

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

The role of web browsing in credit risk prediction

Betty Johanna Garzon Rozo [*], Jonathan Crook, Galina Andreeva

Credit Research Centre, The University of Edinburgh, Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, United Kingdom

ARTICLE INFO

Keywords:
Online retail credit risk
Risk management
Credit scoring
Survival analysis

ABSTRACT

Online mail order and online retail purchases have increased rapidly in recent years worldwide, with Covid-19 forcing almost all non-grocery shopping to move online. These practices have facilitated the availability of new data sources, such as web behavioural variables providing scope for innovation in credit risk analysis and decision practices. This paper examines new web browsing variables and incorporates them into survival analysis as predictors of probability of default (PD). Using a large sample of purchase and repayment credit accounts from a major digital retailer and financial services provider, we show that these new variables enhance the predictive accuracy of probability of default (PD) models at account level. This also holds in the absence of credit bureau data, therefore, the new information can help people who may not have a credit history (thin file) who cannot be assessed using traditional variables. Moreover, we leverage on the dynamic nature of these new web variables and explore their predictive value in short and long- term horizons. By adding macroeconomic variables, the possibility for stress-testing is provided. Our empirical findings provide insights into web browsing behaviour, highlight how the inclusion of non-standard variables can improve credit risk scoring models and lending decisions and may provide a solution to the thin files problem. Our results also suggest a direct value added to the online retail credit industry as firms should leverage the increasing trend of consumers embracing the digital environment.

1. Introduction

The mail order and catalogue sector and many other retailers offer credit together with the goods purchased. Online retail purchases have increased rapidly in many countries over the last decade and especially over the last five years. For example, in the UK online retail sales have growth at a substantially faster rate than in-store sales, increasing from a 3.4% share of all retail sales in 2007 to 27.9% in 2020. And from November 2006 to February 2020, all retailing except automotive fuel online sales had just over a ten-fold increase, showing how online retail sales were already growing strongly prior to the pandemic [1]. In the EU, sales volume by mail order and the internet increased by 54% between 2014 and 2017 (Eurostat, 2017). The recent Covid-19 restrictions made online shopping the preferred and in many cases the only possible way of operation for non-grocery retailers. This rise in online sales on credit enables retail lenders to use new predictors in their credit scoring models that were unavailable before. These new variables describe the way the web is used when making online purchases and characteristics of online access by a credit applicant to their account.

Academic research in credit risk modelling has focussed mainly on which classifier algorithm is on average the most accurate (see [2,3] for reviews) and much less attention has been given to the increase in predictive accuracy that may come from using new types of predictor variables. Recently an emerging literature has moved away from the traditional features (e.g. borrower's application characteristics, credit history, and transactional features) and is exploring "alternative" information. This includes psychometrics [4–10], social network features [11–14], texts characteristics [15–18], mobile phone features [19] and phone usage data [20–24] to predict probability of default (PD). Although, this developing literature (see additional details in Section 2) has demonstrated that "alternative" data offers valuable information to predict loan default, research has focused on applying this type of data mainly in peer to peer (P2P) lending platforms [12–18,21,24,25], credit cards [11,20,22] and traditional bank loans [4,5,7–10]. Very few papers have dealt with credit risk modelling in the online retail credit sector [19,24]. To the best of our knowledge web related variables (e.g. behavioural online characteristics) in the context of the online retail credit industry have not yet been explored in the literature. This paper addresses this gap.

Whilst the alternative variables mentioned above show some

---

[*] Corresponding author.
E-mail address: johanna.garzon@ed.ac.uk (B.J.G. Rozo).

https://doi.org/10.1016/j.dss.2022.113879
Received 22 October 2021; Received in revised form 17 September 2022; Accepted 21 September 2022
Available online 28 September 2022
0167-9236/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

predictability of loan default, their implementation in real online retail credit systems may be impractical for the following reasons. First, some variables such as phone usage data cannot be accessed without explicit permission in developed countries due to data protection regulation laws (e.g. The General Data Protection Regulation 2016/679 –GDPR in Europe, Data Protection Act - DPA in the UK). Second, variables such as social networks simply may not exist (e.g. some people may not have social media accounts). Third, texts and photos are not typically collected by online retailers. Fourth, there may be customer resistance to the use and/or collection of, for example, mobile usage data. Hence, to predict the PD in the context of online retail credit, new data on new variables that are readily available to the service provider, are preferable.

The aim of this paper is to show how web browsing variables, that can be easily collected by online retailers without specifically seeking this additional information from customers, can be incorporated into models of credit risk to predict PD at account level. We aim to show that these new types of variables can enhance the predictive accuracy of credit scoring models in comparison to that achieved by models that include only traditional predictors.

We make several important contributions. We show the impact on predictive accuracy of certain specific web browsing variables in addition to traditional application, behavioural and credit bureau information. Specifically, we first show that for the evaluation of an application for credit to fund online purchases, the inclusion of measures of customer interactions with the online platform including *Number of website visits*, *Number of account sessions*, *Number of terms and conditions views* and *Number of mobile devices used,* on average increases the predictive accuracy of credit scoring models. When a variety of costs are applied to the misclassification of applicants, the increase in accuracy and benefits to the lender are notable, even though the improvement in rank ordering as indicated by area under the Receiver Operating Characteristics curve (ROC) is modest. The predictive contribution of these variables has not been shown before in the literature. Second, we demonstrate the value of web browsing behaviour as a PD predictor variable. We do this by exploring its predictive value over two different time horizons. We show that the inclusion of web browsing variables improves predictive performance over the longer term (a 12 month horizon), but not over a shorter term (a 3 month horizon). Third, we show which new predictors are statistically significant in the new models. This gives confidence in the relationships identified.

These findings suggest it is beneficial for online retailers to collect this information to enhance the accuracy of their credit risk models. Our investigation of the predictive accuracy of a risk model when the new variables are included instead of behavioural variables answers the question as to whether the new information may help people who may not have a credit history or who have a thin file[1] and so no credit score using conventional variables. Enabling such people to gain a credit score including the new variables may facilitate financial inclusion.

The structure of the paper is as follows. Section 2 gives an overview of the literature on new features in credit risk models. Section 3 introduces survival models, the notation used in the paper, and describes the data and experimental set up. Section 4 discusses the results and section 5 concludes.

2. Literature review

Our paper contributes to the recent and increasing literature that has considered the predictive accuracy of additional and new covariates in credit risk scoring models. A number of papers have assessed the characteristics of text used to describe the borrower and loan purpose by P2P applicants as predictors of PD. For example, Gao et al. [15] considered the readability, tone and deceptive cues of text used. Netzer et al. [16] considered two word combinations and word classes. Iyer et al. [17] considered the number of words and whether the applicant included a picture. Dorfleitner et al. [18] examined spelling errors, keywords and the number of words. Using characteristics extracted from photos of both lenders and borrowers in the P2P lending platform, Gonzalez and Loureiro [25] analysed the probability of loan application success.

Other papers have considered psychometric information [4–10] to predict PD. Further studies have considered Facebook likes [12] or characteristics of friendship groups [13]. Lin et al. [13] found that the online friendships of borrowers increase the probability of successful funding. A few papers have considered characteristics of mobile phone usage data such as phone calls, messages, data volume, and app usage [20–24,26] in predicting PD. Social media usage was considered by Lu [26] and social media network size and social media messaging activity was used by Ge [14] and found to increase predictive accuracy.

Several papers have related some aspect of loan performance to types of products purchased using the credit. Wu [27] showed that inclusion of such variables increased the accuracy of a PD model for P2P loans. Vissing-Johansen [28] found that type of product affects the proportion of a loan that is not repaid, but did not relate it to PD. Li [29] showed that the type of product (health or education products) were statistically significant in a PD model but their contribution to predictive accuracy was not shown.

The closest to our paper is by Berg [19]. Berg et al. [19] considered factors affecting PD for an online German e-commerce company, which they described as a “digital footprint”. This included factors such as the type of device used, the operating system, the channel from which the customer comes to the website, the time of day of purchase, the email host and the email provider, checkout time from the website and whether the customer allowed data tracking. Wu [27] considered different types of web based behaviour compared to our paper. Wu included topics of customer interest (health or digital) gleaned from their web usage and whether shopping or photo apps were downloaded onto their device.

However, despite the growing desirability of incorporating Know Your Customer, we cannot find any papers that have considered certain statistical aspects of web usage, apart from Wu [27] who considered search frequency, or of online account access as predictors of the probability of loan default. This is surprising given the rapid increase in using the web to make purchases. Mail order companies and retailers that provide credit online and who ignore the inclusion of new web related variables might see their models become less accurate as online purchases increase. Our paper adds to this literature by quantifying the increase in the accuracy in estimating the probability of a borrower defaulting when new web behaviour variables are included. Our research becomes even more important in the current situation with a rapidly increasing proportion of shopping moving online.

3. Methodology and empirical analysis

3.1. Survival analysis

Credit risk analysis is an essential tool to estimate the probability of a borrower defaulting [3,30] and allows banks and retail lenders to predict the credit risk in their portfolios of either traditional credit products or those offered online. Traditional scorecards predict the PD in a given time period. However, these fixed-period scorecards do not consider time-dependent characteristics. This difficulty can be overcome by implementing survival analysis [31–33]. Survival analysis (SA) models the probability that a credit account will remain in a particular state (for example that payments are up to date) until a chosen time when it will move into a different state [3]. For example, it models the probability that an account will move from being up to date with payments to being in default, for the first time, within a time period of interest, e.g. 6

---

[1] A “thin file” refers to the credit report of someone with little or no credit history. A bank or lender may be unable to calculate a credit score since there is not enough information in the credit history to do so.

months or 12 months. A survival model differs from a cross-sectional model in that instead of modelling whether the customer defaults in a fixed period it models *when* the event (e.g. default) will occur [34]. The advantages provided by survival analysis are discussed in [31–33,35–38], among others.

3.2. Discrete time

In this paper we follow the literature [37,39,40] and treat time as discrete since data are observed monthly. We assume a discrete time hazard function with a pre-specified, but very flexible baseline hazard function. Thus, we use the following model

$$log\left(\frac{P_{i,t}}{1-P_{it}}\right) = g(t_m) + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{\omega}_{it-l} + \boldsymbol{\beta}_3^{\mathrm{T}}\mathbf{z}_{t-l}, \tag{1}$$

where

- P_{it} denotes the probability of default for an account i in a (discrete) duration period t, $t = 1, 2, \ldots, \mathrm{T}$;
- $\mathbf{x}_i$ denotes a column vector of covariates whose values are specific to a borrower, i, and do not vary over time (application variables);
- $\boldsymbol{\omega}_{it}$ denotes a column vector of covariates whose values differ between borrowers and differ over time (behavioural and transactions variables);
- $\mathbf{z}_t$ denotes a column vector of covariates that vary over time but not between individuals (macroeconomic variables);
- $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ denote column vectors of coefficients to be estimated;
- $g(t_m)$ is the baseline hazard function; and
- l denotes a lag.

The functional form of the baseline, $g(t_m)$, must be specified. Various options are available [31,37–39]. In this paper, we use spline functions, specifically cubic spline basis[2] functions, because they are very flexible for the $g(t_m)$ function. Notice also that macroeconomic variables are measured in calendar time and for each case their value must be matched to duration time using the relationship $t = o + c$ where o denotes calendar time of account origination and c is calendar time. We estimate the parameters of eq. (1) using pooled logistic regression estimators. To ensure (*a*) a hazard model is estimated rather than a cross-sectional PD model and (*b*) the model represents the probability that default happens for the first time in period *t*, the data matrix used in the estimation must be set up in a specific way. The dependent variable, the indicator of default, is coded 0 for all time periods before the period of default, 1 in the month of default and missing thereafter. See [37,38,43,44].

3.3. Data

Our dataset consists of active accounts that were opened between January 2013 and January 2017 and provided by a major UK online retailer and financial services provider. Performances were observed until October 2017. We have access to purchase and repayment information for this sample. The purchase process is as follows. A potential purchaser visits the company's website to search for the product the client wishes to buy. If the customer decides to buy a product, the purchaser will be given a choice between paying within a fixed period or repaying the cost of the product over 6 or 12 months after the order is placed, with the interest added to the price from the date of the order. If a purchaser has not purchased a product from this company before, the customer would complete an application form for an account. The potential purchaser (i.e. applicant) is credit scored using an application-scoring model and if the applicant score exceeds a threshold, the applicant is granted credit up to a specific limit and the product is delivered to the new client. If the purchaser already has an open account, the purchaser is still credit scored using a behavioural model. If that score exceeds a threshold and if the price of the product, when added to any outstanding balance, does not exceed the credit limit then the purchaser is granted the credit. The purchaser can then choose between paying the full price immediately or taking credit and the product is delivered to the customer. In essence, if a customer has an open account this is a credit line to which the customer can add further credit up to a set limit. Therefore, a survival analysis is needed because our data has time varying behavioural variables and PD values in any chosen future period can be predicted.

3.4. New web behavioural variables

We have available a number of traditional characteristics of each applicant, for example age, socio-demographic category, bureau variables[3] and a number of behavioural variables. In addition we have access to a number of additional variables that are not commonly used in credit scoring models and that we wish to assess for their predictive performance. These are listed in Table 1 where we give the definition and variable name used in the analysis, for each variable. We call these variables "new behavioural web related variables". These variables are related to the interaction between the customer and the retailer's website.

We are unable to give summary statistics for each variable because of commercial confidentiality. We might expect that several of these new behavioural web related variables would be predictive of default. Hence we might expect that customers who were concerned about their ability to repay their balance might look at their balance, which includes accrued interest, more frequently than customers who were well able to repay because they may be repeatedly computing how much interest they would incur relative to the benefits gained by making further expenditures or debt repayments instead.

Previous research has also provided some evidence that psychological traits are related to credit performance; in particular, neuroticism or anxiety can be positively related to the number of missed credit repayments [45]. It is logical to expect that people who are more anxious visit the website more often and show more erratic behaviour with a higher number of visits than less anxious people do. In terms of devices used, one might also argue that the number of devices used to access the company website, proxies for income.

To the best of our knowledge, the new behavioural web related variables relating to web browsing that we have access to are novel and have not been considered in traditional credit risk models literature for predicting PD. The inclusion of these new variables would potentially generate more accurate predictions of default or risk scores because they take a more complete view of risk. Another advantage of including such new variables into credit risk modelling is that it enables the PD to be computed using the most recent (behavioural) data available.

3.5. Other variables

Transactions variables are also available for inclusion. These are listed in Table 2. We also consider for inclusion several macroeconomic

---

[2] A cubic spline basis function of a variable consists of a separate cubic polynomial within each of several selected ranges of values of that variable where the separating values between the ranges are known as knots. So, for example, a cubic spline function with four knots, γ_1, γ_2, γ_3 and γ_4, can be written as $f(X) = \gamma_0 + \gamma_1 X + \gamma_2 X^2 + \gamma_3 X^3 + \gamma_4 (X - k_1)^3 + \gamma_5 (X - k_2)^3 + \gamma_6 (X - k_3)^3 + \gamma_7 (X - k_4)^3$ or more generally $f(X) = \sum_{j=1}^{j+d} \gamma_j B_j(X)$ where d is the power of the polynomial and B_j is a set of basis functions and j is knot j. See [41; 42].

[3] Bureau variables are predictive variables collected by a credit bureau that gathers and researches individual credit information from various creditors and sells that information to private lenders for a fee. Thus, these variables help lenders decide whether to grant credit to a new applicant.

Table 1
New Behavioural web related variable names and definitions available for selection.

Panel (a) Variables where summary statistics were available

Variable	Definition	Average in month t	Max in month t	Median over account life	Mean over account life
Number of devices used	Number of devices used per customer in month t	AveragеperMonth_num_device	Max_num_device	Median_num_device	Mean_num_device
Number of my account sessions	Number of logons a customer made into his/her account in month t	AverageperMonth_myaccount	Max_num_myaccount	Median_num_myaccount	Mean_num_myaccount
Number of payment sessions	Number of payments made per session in month t	AverageperMonth_num_payment	Max_num_payment	Median_num_payment	Mean_num_payment
Number of terms sessions	Number of times a customer reviews his/her terms and conditions in month t	AverageperMonth_num_terms	Max_num_terms	Median_num_terms	Mean_num_terms
Number of website visits	Number of visits a customer made to the retailer's website in month t	AverageperMonth_ website	Max_website	Median_website	Mean_website

Panel (b) Variables where raw values in month t were available

Number of:	Value of:
Clothing purchases	Clothing purchases
Electrical purchases	Electrical purchases
Furniture purchases	Furniture purchases
Purchases of other goods	Purchases of other goods
Purchases in total	Purchases in total
Seasonal items	Seasonal items
Returned items	Returned items
Rejected items due to insufficient credit	Rejected items due to insufficient credit

Table 2
Transactional variables available for selection.

Variable	Definition	Name
Total outstanding balance	Balance on account at closing date in month t.	Total_outstanding_balance
Balance divided by credit limit	Total_outstanding_balance/ Credit_limit.	Ratio_Balance_vs_CreditLimit
Buy now pay later balance	Buy now pay later balance at closing date in month t.	Bnpl_Balance
Total balance divided by bnpl balance	Total_outstanding_balance/ Bnpl_balance.	Ratio_Balance
Credit limit in month t relative to credit limit in the previous month	Credit_limit in month t/Credit limit in month t-1.	Ratio_creditlimit
Value of scheduled payment	Value of scheduled payment from previous statement.	Sched_pay_prev_stat
Full payment of balance before end of delayed payment period	Full payment of the balance account before delayed payment period ends.	Paym_off_account
Cumulated arrears	Amount accrued from the date on which the first missed payment was due.	Arrears_amount

variables which are described in Table 3. We chose these variables because we expect them to be correlated with PD and because there is evidence in the literature that suggests they are [31,37,39,46–49]. For example, one might expect that if the unemployment rate or bank rate are high then on average PD would be high as people have less disposable income from which to repay loans. The house price index and FTSE both represent wealth of different kinds which would be expected to be negatively correlated with PD and the index of production is a proxy for average income.

Table 3
Macroeconomic variables.

Variable	Description	Source
UER	Unemployment rate (aged 16 and over, seasonally adjusted)	ONS
IP	Index of Production in the UK. Seasonally adjusted, January 2008 to March 2018. Source: Primarily Monthly Business Survey (Production and Services)	ONS
BR	Official Bank Rate Bank of England	BOE
IUM2WDT	Monthly interest rate of UK monetary financial institutions (excl. Central Bank) sterling 2 year variable rate mortgage (95% LTV) to households (in percent) not seasonally adjusted	BOE
IUMB479	Monthly interest rate of UK monetary financial institutions (excl. Central Bank) sterling 2 year variable rate mortgage (90% LTV) to households (in percent) not seasonally adjusted	BOE
IUMCCTL	Monthly interest rate of UK monetary financial institutions (excl. Central Bank) sterling credit card lending to households (in percent) not seasonally adjusted	BOE
LPMB4TF	Monthly amounts outstanding of other consumer credit lenders (excluding the Student Loans Company) sterling consumer credit lending to individuals (in sterling millions) not seasonally adjusted	BOE
HOUSE PI	House price index	NBS
FTSE100	FTSE 100 monthly average (daily data). It is an index composed of the 100 largest (by market capitalisation) companies listed on the LSE	LSE
RPI	The retail prices index or retail price index (RPI) is published monthly by the Office for National Statistics. We took RPI All Items Index: Jan 1987 = 100	ONS
GfK	Consumer confidence index is an economic indicator that measures the degree of optimism that consumers feel about the overall state of the economy and their personal financial situation	GfK

ONS = Office for National Statistics. BOE = Bank of England, NBS=Nationwide Building Society, LSE = London Stock Exchange. GfK = Growth from Knowledge is a global consulting service for the consumer products industry.

3.6. Variable selection and model set-up

We coarse classified both categorical and numeric input variables and represented their original values as weights of evidence.[4] Both are commonly implemented in practice [3,34,50]. The variables that were eventually included in the models were the result of a two stage selection procedure. The first stage consisted in selecting variables individually using pre-screening methods (based on Gini and Information Value) prior to running the survival analysis. From the variables that passed the first stage of selection a second stage consisted of a stepwise survival model.

We carried out this procedure for covariates where the time-varying covariates were lagged 3 months and, separately, where they were lagged 12 months. That is, in eq. 1, ι, the lag length is taken to be either 3 months or 12 months.

Whilst the predictive horizon could be set at any time, we illustrate the predictive enhancement within in a 12 months period and within a 3 months period. Twelve months is the standard time horizon for PD estimation for regulatory models such as IFRS9, and a 3 month period that is a common period of prediction among practitioners. The final variables are shown in Table 4.[5]

The sample was split into two non-overlapping independent data sets; a modelling dataset for estimating and validating the model and a second dataset for testing (i.e. out-of-time). The modelling data was split randomly with 70% of accounts selected for training and the remaining 30% selected for validation. The models were parameterised using the training data and then assessed on the validation data set. The new behavioural web related variables for each account in the modelling data were observed until the account defaults for the first time or until the end of the observation time window. An account was considered in default if it had missed three consecutive payments. The new behavioural web related variables and the macroeconomic variables were lagged either by 3 months or by 12 months and we estimated separate models for each lag length. Thus a model with covariates lagged 3 months (12 months) yields predictions 3 months (12 months) into the future. For convenience we denote models with covariates lagged 3 months (12 months) as 'Lag 3' ('Lag 12'). The modelling data set for models with Lag12 correspond to accounts opened from January 2013 to December 2015 with performance observed until December 2016. The testing data set for models with Lag12 consists of accounts opened between January 2016 and October 2016 and that did not default before October 2016. Their performance was observed until October 2017.

For each Lag12 test account a prediction of the probability that the account defaulted in *any month* between November 2016 and October 2017 was computed. These predictions were made by substituting the values of the covariates at October 2016 into the parameterised eq. 1 to predict the survival probability over the 12 month period where the change in the future probability between months was due to the change in the duration time variable only. Thus, the testing data is out-of-sample and (largely) out-of-time relative to the training data set whereas the validation data is out-of-sample but in-time. We argue that the use of October 2016 values of the covariates is justified because practitioners would typically use the latest values of the covariates that would be available to them.

Turning to models with Lag3, the modelling data set corresponds to accounts opened from January 2013 to December 2015, with performance observed until April 2016. The test data set for models with Lag3 consists of accounts opened from January 2016 to January 2017, provided they did not default before end January 2017. Their performance was observed until April 2017. For each test account with Lag3 a prediction of the probability that the account defaulted in *any month* between February and April 2017 was computed using the same procedure as for lag12. The sample sizes of the training, validation and test samples for models with each lag length are listed in Table 5.

We constructed four survival models with application, transactional, bureau and new behavioural web related variables for models with Lag3 and we used different variables for models with Lag12 depending on which were selected by the two stage selection process described above. Details of the variables were shown in Tables 1, 2, and 3. We chose four combinations of these variables for each lag as detailed in Table 6 to assess the contribution to predictive accuracy of the new behavioural web related variables.

All four models include application, transactional and macroeconomic variables. Model A1 corresponds to models where bureau variables were also included but no new behavioural web related variables. A2 includes bureau and the new web variables. A3 does not include either bureau or the new behavioural web related variables and A4 adds only the new web variables.[6] To assess the performance of the new behavioural web related variables we compare the performance of model A1 with that of model A2, where the same application, transactional and macroeconomic variables are included, the only difference being the additional of the new web variables in A2 but not in A1. We also compare the performance of model A4 (which includes the new variables) with that of A3 that omits the new variables, in these two cases bureau variables are omitted. The latter comparison allows us to consider the performance of the web variables for applicants that do not have a previous credit history.

3.7. Assessing performance

We compare the predictive performance of the competing models using three criteria. Firstly, we use two standard statistical measures, Kolmogorov-Smirnov statistics (KS) and Receiver Operating Characteristic curves (ROC), that are very commonly used in the credit scoring literature [2,3,51–53]. Secondly, as is commonly used in the literature (see above references) and by practitioners, we also use metrics that require an account to be predicted to be either good or bad: accuracy and sensitivity (the proportion of bad cases predicted to be bad). The prediction is made by comparing the predicted survival probability over a given time horizon with a cut-off probability. The cut-off is computed from the training data set for each model such that the proportion of cases observed to be good over the horizon equals the proportion of cases predicted to be good over the same horizon. Cases are ranked in the training sample in ascending order by predicted hazard value (i.e. PD). The case located at the percentile equal to the proportion of observed Good cases is taken and its value of predicted hazard is the cut-off threshold. Then this cut-off is applied to the out-of-time test set to make predictions. Thus, the out-of-time test can be split into two sets, Good and Bad cases, and defining those above the cut-off as Bads and

---

[4] We create bins in order to reduce the number of unique levels. We assessed the impact of the binned variables on the model by checking the sensitivity of the model with and without the binned variables. We found that models using binning attributes lead to stronger predictive power than those that do not use binned variables. Therefore we bin all our predictors into 5 groups, each with an equal number of observations, and 'missing' as a separate level. We interactively split and combine the initial bins, creating classes to attempt to improve the predictive power of each input.

[5] We are unable to provide details for the bureau variables used due to commercial confidentiality; but typical generic examples are: number and types of credit accounts, how long each account has been open, amounts owed, amount of available credit used, whether bills were paid on time, number of recent credit inquiries, percentage of houses at the postal code with a registered default, etc.

[6] Each model in Table 6 was trained with two types of values for the independent variables. First, we use dummy variables, i.e. categorical effects from the coarse classification. Second, we use weights of evidence of the coarse classes. Since we found that the second one provided slightly better results, we only present the results for the models that use weights of evidence of the coarse classes as the values for the covariates.

Table 4
Covariates included in each set of models after the selection procedure was implemented.

Models Lag12

Application	Transaction	Macroeconomic	Bureau	New web behavioural
Socio-economic_segment3	Ratio_Balance_vsCreditLimit	FTSE100	Bureau1	AverageperMonth_num_device
Brand_type	Ratio_Balance	GfK	Bureau2	Max_num_device
Customer_age	Ratio_Creditlimit	IP	Bureau3	Max_num_payment
	Sched_pay_prev_stat		Bureau4	Max_num_terms
	Paym_off_account		Bureau5	Max_website
	Arrears_amount		Bureau6	Mean_num_devices
			Bureau7	Mean_num_payment
			Bureau8	Mean_website
			Bureau9	Median_num_payment
			Bureau10	Number_purchases
			Bureau11	Value_electrical
				Value_rejected_items

Models Lag3

Application	Transaction	Macroeconomic	Bureau	New web behavioural
APR	Ratio_Balance_vs_CreditLimit	FTSE100	Bureau1	AverageperMonth_num_device
Socio-economic_segment2	Ratio_Balance	GfK	Bureau2	Max_num_device
Brand_type	Ratio_Creditlimit	HousePI	Bureau3	Max_num_payment
Customer_age	Sched_pay_prev_stat	IP	Bureau4	Mean_num_of_myaccount_sessions
		IUMB479	Bureau5	Mean_num_payment
		IUMCCTL	Bureau6	Number_of_reject_items
		LPMB4TF	Bureau7	Number_returned
		RPI	Bureau8	Value_rejected_items
		UER	Bureau9	Number_website
			Bureau10	
			Bureau11	
			Bureau12	
			Bureau13	
			Bureau14	
			Bureau15	

Table 5
Sample sizes.

Model	Training and validation data	Number of accounts	Test data	Number of accounts	Total accounts
Lag12	Opened from Jan 2013 to Dec 2015	26,270	Opened from Jan 2016 to Oct 2016	6,050	32,320
Lag3	Opened from Jan 2013 to Dec 2015	26,270	Opened from Jan 2016 to Jan 2017	9,267	35,537

Table 6
List of models.

Model	Application + Transactional + Macroeconomic variables	Bureau variables	New web related behavioural variables
A1	Yes	Yes	No
A2	Yes	Yes	Yes
A3	Yes	No	No
A4	Yes	No	Yes

those below it as Goods. A confusion matrix is created for the out-of-time test using a different cut-off for each model. Advantages of this way of setting the cut-off are that it is independent of the model estimated (since the observed proportions of goods and bads are independent of the model). It is also independent of the corresponding holdout sample.

Thirdly, we compare the models in terms of misclassification costs. It is well known [3,31,54,55] that the cost of mis-classifying a Good account as a Bad, equal to the opportunity cost of lost interest [3] is smaller than that resulting from mis-classifying a Bad account as a Good, when the institution loses some or all of not only the interest but also the repayment of principal [56]. Therefore, to calculate the relative misclassification cost, we apply a cost function that penalises the Type II error (observed Bad predicted as Good) as follows: (1) a correctly classified case has no cost (cost = 0), (2) a Good case predicted as Bad has a cost of 1 and (3) a Bad case wrongly predicted as Good incurs a cost of 20. Observed values of these cost ratios are not in the literature, but these relative cost penalties have been used before in the literature [31] because they are believed to be realistic. To demonstrate robustness, relative costs of 15 and 25 are also reported for both sets of models, Lag 3 and Lag 12.

4. Results and findings

4.1. Baseline survival function

Fig. 1 shows the empirical baseline Survival function (left hand scale) and the Hazard function (right hand scale) for the training data across the full time period of 48 months. The horizontal axis is the duration time since account opening. The survival plot (blue line), shows a typical decline over time that is consistent with the literature [36]. From the hazard function (red line) we can see that the hazard of default is higher between the months 4 and 9 and then decreases drastically, especially between the months 15 and 30. After month 30, the PD has a constant mean but is quite changeable probably due to the relatively few cases in this region.

4.2. Model assessment and predictive performance

Table 7 shows results for the models with Lag 12. Panel (a) shows a comparison of the performances of the estimated models using proportions of cases and panel (b) shows the predicted costs of bad cases

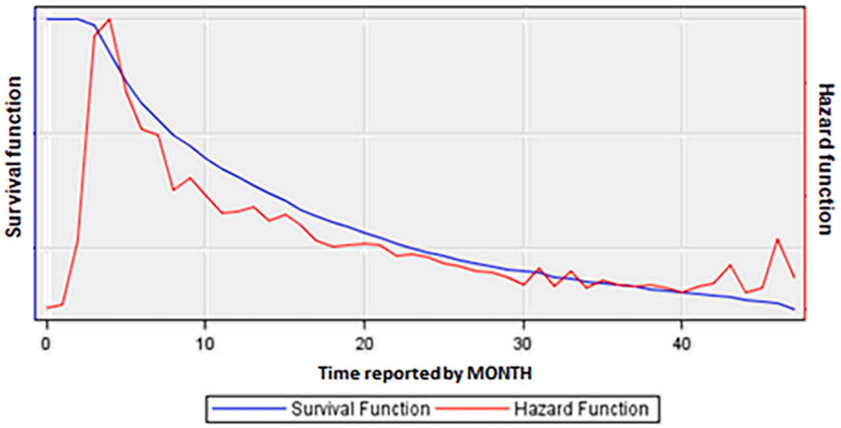

Fig. 1. Baseline survival curve.
Note. The units on both axes have been removed for commercial confidentiality reasons.

Table 7
Models with Lag12.

Panel (a) KS and ROC measures

		Models with Lag12			
		A1	A2	A3	A4
	Input variables	23	35	12	24
	Significant variables	23	35	12	24
Sensitivity	Test (out-of-time)	0.8333	0.8674	0.7412	0.7942
	Train	0.4670	0.4673	0.4273	0.4320
KS	Validation (in-time test)	0.4533	0.4600	0.4099	0.4132
	Test (out-of-time test)	0.4360	0.4500	0.3590	0.3710
	Train	0.7781	0.7807	0.7526	0.7572
ROC	Validation (in-time test)	0.7706	0.7743	0.7491	0.7550
	Test (out-of-time test)	0.7908	0.7981	0.7376	0.7431

Panel (b) Misclassification costs for out-of-time data set

	Models with Lag12			
Cost on bad cases	A1	A2	A3	A4
15	3,868	3,710	4,722	4,521
20	4,528	4,235	5,747	5,336
25	5,188	4,760	6,772	6,151

Table 8
Models with Lag3.

Panel (a) KS and ROC measures

		Models with Lag3			
		A1	A2	A3	A4
	Input variables	32	41	17	26
	Significant variables	25	33	16	23
Sensitivity	Test (out-of-time)	0.8244	0.8197	0.8478	0.8009
	Train	0.5490	0.5397	0.5215	0.4850
KS	Validation (in-time test)	0.5946	0.6019	0.5297	0.5348
	Test (out-of-time test)	0.6340	0.6280	0.6570	0.6220
	Train	0.8245	0.8254	0.8081	0.8004
ROC	Validation (in-time test)	0.8533	0.8570	0.8109	0.8234
	Test (out-of-time test)	0.8865	0.8880	0.8910	0.8800

Panel (b) Misclassification costs for out-of-time data set

	Models with Lag3			
Cost on bad cases	A1	A2	A3	A4
15	3,050	3,023	2,884	3,014
20	3,425	3,408	3,209	3,439
25	3,800	3,793	3,534	3,864

misclassified in the out-of-time data set, when the optimal cut-off is computed from the training data set. The same results are presented for Models with Lag 3 in Table 8 panel (a) and panel (b). The row '*Input Variables*' shows the number of variables selected from those in Table 4 using the selection procedure explained in section 3.5. The row '*Significant Variables*' gives the number of significant variables in the model. We first consider models with Lag 12. The effect of including the behavioural web related variables is indicated by the performance of model A2 relative to that of A1 since both have the same application, transactional and macroeconomic variables, but whilst A2 includes the new variables A1 does not. Although the performance uplift from A1 to A2 is modest in terms of ROC and KS, in terms of sensitivity and minimising misclassification costs A2 is clearly superior. Model A2 classifies 86.74% of bad cases correctly whereas model A1 correctly classifies 83.33%.

The relative cost advantage of A2 versus A1 is most evident for a relative cost of 20 or higher. For instance, the cost reduction gain by model A2 when comparing it with model A1 is 6.5% and 8.3% at cost ratios 20 and 25 respectively.

If we compare the models with and without the new variables but without the bureau variables, we see that A4 (with the new variables) has higher ROC, KS and sensitivity values than A3 (without the new variables). Again, the model with the new variables has lower misclassification costs compared with the model without them; in this case a reduction in cost of 7.2% and 9.2% at cost ratios 20 and 25 respectively when the new web related variables are included. In general, based on the KS and areas under the ROC for the test data, the most predictive model for models with Lag 12 is model A2, which incorporates application, transaction, macroeconomic, bureau and the new behavioural web related variables. This is also confirmed by the graph of the ROC curves in Fig. 2. The sensitivities yield the same conclusion. Notice that the results from Table 7 panel (a) are consistent for both KS and ROC statistics and across all three data sets: training, validation (in-time) and test (out-of-time or scoring) data sets.

Turning to the results for models Lag 3, from Table 8 panels (a) and (b) we see that web- based variables do not enhance predictive accuracy when only a 3 month prediction is required. Thus, we conclude that web browsing data enhances the predictive accuracy in the long-term but not in the short term. We argue that this better performance for a 12 month horizon is particularly useful since the Basel Accord, see BIS 2015 [58], requires PD to be predicted over this longer horizon and because over a short term period a particular borrower's behavioural pattern may be merely temporary or circumstantial.

While the results for models Lag 12 are encouraging, additional robustness check were conducted to support these results. As a robustness check, we conduct a five-fold Cross-validation [59,60] to evaluate

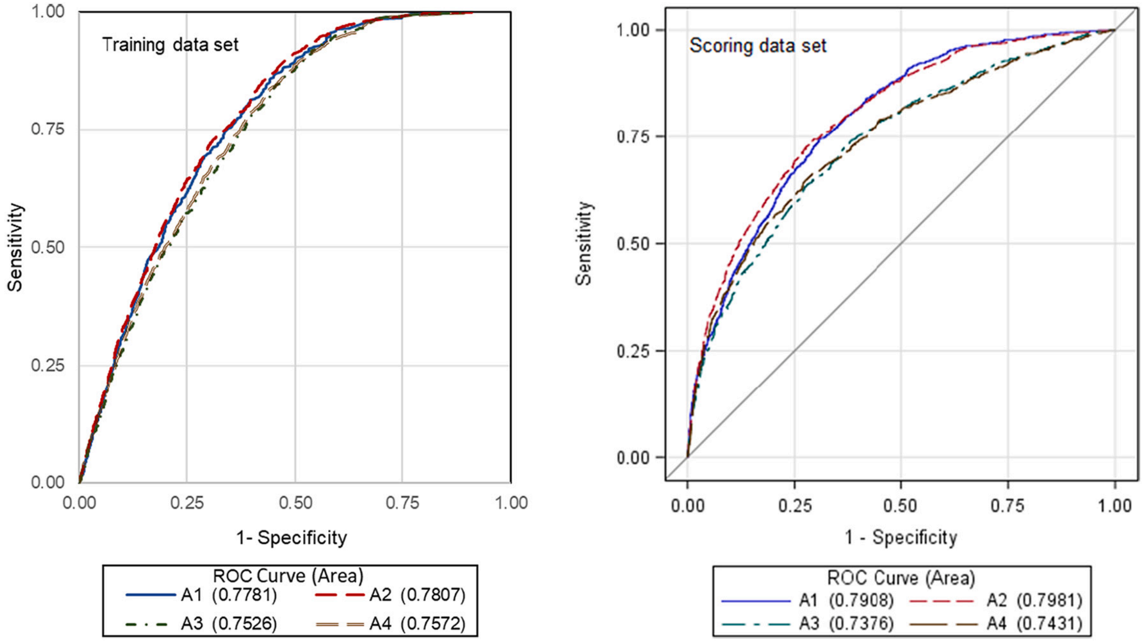

Fig. 2. ROC curves Lag 12 model A2.

the predictive power of the models. We randomly split the modelling sample into 5 equal-sized sub-samples. We train the model on four of the five sub-samples and test it on the sub-sample left out. The test sample in this case is out-of-sample but in-time. We repeat this procedure so that each sub-sample is left out once. We then repeat this procedure using a different seed value for generating the random partitions. Thus, we train 10 models for each set of variables in A1, A2, A3 and A4. We do this for models with Lag 12 only because that is the time horizon over which the web variables enhance predictive accuracy. We estimate the mean of four statistics: sensitivity, KS, ROC, and misclassification costs.[7] Table 9 summarises the Cross-validation results. The results in Table 9 verify the conclusions from Table 7 that model A2 gives greater predictive accuracy than A1 and A4 gives greater accuracy compared with A3. In fact, the same variables are statistically significant in the Cross-validation models as in the single partition models (Table 7). As the results reveal, based on the average of the KS, areas under the ROC, sensitivity and misclassification costs for the validation and test data sets (Table 9) and on the results for the test data in the single partition (Table 7), model A2 shows the highest prediction accuracy for models with Lag 12.

Table 9

Cross-validation results for Models lag 12.

		A1	A2	A3	A4
	Input variables	22	34	12	24
	Significant variables	21	33	11	23
Sensitivity	Test (out-of-time)	0.7997	0.8373	0.7848	0.8199
KS	Train	0.4577	0.4623	0.4124	0.4213
	Validation (in-time test)	0.4623	0.4667	0.423	0.4294
ROC	Train	0.7717	0.7767	0.7496	0.7557
	Validation (in-time test)	0.7710	0.7749	0.7485	0.7533
Misclassification Costs (out-of-time)	Cost on bad cases 15	4,247	4,013	4,584	4,484
	Cost on bad cases 20	5,040	4,658	5,436	5,197
	Cost on bad cases 25	5,833	5,302	6,289	5,910

Finally, we compute paired *t*-tests to assess the statistical significance of differences in model performance for each statistic. Table 10 shows these results.

A paired *t*-test is the most relevant in our context since we have pairs of measurements for each model that were obtained from the same sample. An overview of learning algorithms evaluation and a description of selected statistical significance tests can be found in [60]. From Table 10 we can safely reject the null hypothesis, H_0 (i.e. the mean difference between pairs of measurements is zero), at the 0.05 and 0.01 level of significance for model A1 and A2 regarding ROC, sensitivity and misclassification costs for the three scenarios (i.e. 15, 20, 25). We can conclude that model A2 and A1 show a statistically significant performance difference on the validation and test data sets. A similar result is observed for models A3 and A4. Overall, A4 and A3 also show a statistically significant performance difference on ROC, sensitivity and misclassification costs for two scenarios (i.e. 20 and 25) in the validation and test data sets. However, the observed differences between pairs of measurements for KS for the models A1 and A2 and between A3 and A4 are not significant at 5%. Hence, the results strongly suggest that model A2 performs better than A1 and model A4 performs better than A3 since they have greater ROC, sensitivity and lower misclassification costs.

Notice also that the difference between pairs of measurements for ROC, sensitivity and misclassification costs are statistically significant for differences between A2 and all the other models. Though KS is not statistically significantly different between A1 and A2, and A3 and A4, the incorporation of web behaviour variables provides relevant information to assess PD at account level and models that include them outperform models that omit them. In addition, we verify the three

---

[7] The individual results for the 40 models are available on request.

Table 10

Paired *t*-test statistics value for Models lag 12.

		A2		A3		A4	
		t-value	*p-value*	*t-value*	*p-value*	*t-value*	*p-value*
Sensitivity	A1	−7.94	<0.0001	4.30	0.0026	−4.39	0.0023
Test (out-of-time test)	A2			14.32	<0.0001	7.04	0.0001
	A3					−11.76	<0.0001
KS	A1	−1.56	0.1531	10.28	<0.0001	11.17	<0.0001
Validation (in-time test)	A2			9.09	<0.0001	10.18	<0.0001
	A3					−2.57	0.0304
ROC	A1	−5.45	0.0004	22.18	<0.0001	18.99	<0.0001
Validation (in-time test)	A2			18.05	<0.0001	23.56	<0.0001
	A3					−5.39	0.0004
Misclassification	A1	5.28	0.0007	−13.72	<0.0001	−4.78	0.0014
Costs on bad cases 15	A2			−15.59	<0.0001	−13.08	<0.0001
Test (out-of-time test)	A3					2.13	0.0655
Misclassification	A1	6.18	0.0003	−10.57	<0.0001	−2.52	0.0356
Costs on bad cases 20	A2			−15.40	<0.0001	−12.42	<0.0001
Test (out-of-time test)	A3					4.23	0.0029
Misclassification	A1	6.64	0.0002	−8.96	<0.0001	−1	0.3474
Costs on bad cases 25	A2			−15.23	<0.0001	−11.78	<0.0001
Test (out-of-time test)	A3					5.65	0.0005

Table 11

Coefficient estimates from Model Lag 12 A2.

Category vars.	Effect	Coefficient	Odds RE	Wald CS	*p-value*
Application	Soc Econ Segment3_w	−0.653	0.521	178.460	<0.0001
	Brand_type_w	−0.664	0.515	180.433	<0.0001
	Customer_age_w	−0.242	0.785	14.599	0.0001
Transactional	Ratio_Balance_vsCreditLimit_w	−0.562	0.570	69.072	<0.0001
	Ratio_Balance_w	1.675	5.340	46.432	<0.0001
	Ratio_Creditlimit_w	1.281	3.600	40.254	<0.0001
	Arrears_amount_w	1.206	3.339	14.652	0.0001
	Sched_pay_prev_stat_w	−0.439	0.645	15.337	<0.0001
	Paym_off_account_w	2.603	13.508	18.873	<0.0001
Macro-economic	FTSE100_w	0.355	1.425	6.595	0.0102
	GfK_w	0.737	2.089	137.099	<0.0001
	IP_w	−0.189	0.828	6.709	0.0096
Bureau	Bureau1_w	0.336	1.399	12.607	0.0004
	Bureau2 _w	−2.154	0.116	13.902	0.0002
	Bureau3_w	−0.449	0.638	25.353	<0.0001
	Bureau11_w	−0.204	0.815	4.644	0.0312
	Bureau4_w	0.871	2.388	25.532	<0.0001
	Bureau5_w	−0.316	0.729	10.040	0.0015
	Bureau6_w	−0.408	0.665	22.821	<0.0001
	Bureau7_w	0.958	2.605	8.124	0.0044
	Bureau9_w	−0.483	0.617	59.637	<0.0001
	Bureau10_w	−0.671	0.511	302.308	<0.0001
	Bureau8_w	−0.503	0.605	77.521	<0.0001
Web behavioural	AverageperMonth_num_device_w	−2.215	0.109	63.641	<0.0001
	Max_num_device_w	−1.320	0.267	20.632	<0.0001
	Max_num_payment_w	−2.144	0.117	67.224	<0.0001
	Max_num_terms_w	−0.552	0.576	8.818	0.003
	Max_website_w	2.078	7.985	39.653	<0.0001
	Mean_num_devices_w	−0.948	0.387	32.857	<0.0001
	Mean_num_payment_w	2.070	7.927	39.498	<0.0001
	Mean_website_w	−1.164	0.312	52.513	<0.0001
	Median_num_payment_w	3.855	47.226	67.672	<0.0001
	Number_purchases_w	−0.803	0.448	8.442	0.0037
	Value_electrical_w	1.892	6.631	32.060	<0.0001
	Value_rejected_items_w	−0.537	0.584	10.047	0.0015

Notes: *Wald CS = Wald Chi-Square; Odds RE = Odds Ratio Estimate; "_w" stands for weight of evidence.

assumptions[8] of the *t*-test, which must all be verified for the paired *t*-test's results to be valid. In results no presented here we found that all the assumptions of the *t*-test are met.

4.3. Model outputs

This section discusses some of the regression parameters from the most predictive models presented in Table 7, model A2. They are detailed in Table 11. This table shows which new behavioural web related variables are statistically significant and retained by the selection routine.

We notice that variables such as *Median_num_of_paym, Max_num_of_payment* and *AverageperMonth_num_device* are more significant than the other new variables, since they show larger Wald Chi-Square values. Other significant variables are socio-demographic variables (socio-economic segment), consumer confidence (GfK), total outstanding balance/credit limit and a bureau variable.

The variables in Table 11 are weights of evidence which are typically not monotonic with respect to the raw underlying variable they relate to. It would be impractical to present this relationship for each variable and it may breach commercial confidentiality if we did so. However we can consider one variable in detail: Max_website (*Maximum number of website visits)* per month. This example will illustrates the nature of the insights the new web information can provide to lenders.

For all but those aged over 30 years there is a positive relationship between maximum number of website visits and default rate and for younger age groups it is positive when the maximum number of visits per months is 3 or more. This might be explained by a theory of *financial well-being*, where research has shown that there are significant associations between problematic internet use and depression, anxiety and stress [63]. Customers that show emotional instability or higher levels of stress and/or anxiety tend to have excessive internet use [64,65]. Excessive use of the internet may lead to behavioural problems if its use becomes uncontrolled. In addition, unhealthy spending and poor saving behaviour are also correlated with personal stress and anxiety [66]. Hojman et al. [67] found that depressive symptoms are higher for those who have been persistently over-indebted.[9] Our results are consistent with these theories by showing that customers who visit the retailer's website more frequently have a higher PD, as shown in Fig. 3, panel (c) relative to panels (b) and (a).

This observation is more evident in young adults (customers between 18 and 25 years old) than in older groups. Age can also have a significant effect on debt, where younger householders are more likely to be in debt than older householders [68,69]. This could be explained by several reasons, for example it is an individual's lifetime income and consumption profile: the young borrow when expenditure exceeds income due to family commitments; it could also be because of the lack of basic financial knowledge, making young adults take poor financial decisions [70]. The overall default rate for the oldest group (i.e. $customer\_age > 47$) is the lowest of the five groups (panel d). But those in this age group that visit the web most frequently (panel c) have a default rate twice the average for their group (panel d). This rate (in panel c) is close behind that of the youngest adults and almost the same default rate for customers between 25 and 38 years old who visit the web frequently. Increases in the maximum number of website visits per month (*Max_website)* from 3 to 6 visits to over 6 visits (Fig. 3 panel b vs panel c) are associated with increases in the default rate for all age groups. Interestingly, the default rate for heavy website users is relatively independent of Age.

4.4. Models for applicants without credit history

In this section we show that if one omits credit history variables but use web browsing data instead we can gain commercially acceptable levels of predictive performance. This is important because in many countries significant proportions of the over 18 population either have no credit history or have only thin files. Having a thin file is problematic making it difficult to get credit. Demirguc-Kunt et al. [71] estimated almost 2 billion people in the world did not have an account with a financial institution.

We train models with Lag 12, using several combinations of the (weights of evidence of the) following variables: *Number of views of the customer's account, Number of visits to the company's website, Number of terms and conditions checked* and *Number of mobile devices used* only and their derivative variables (i.e. average per month, max, median and mean) plus two application variables (customer age and brand type) and the macroeconomic variables. A five-fold Cross-validation was performed as well to provide robust results (see Section 4.2 to review the procedure). The results for the model with the highest ROC are shown in Table 12 and Table 13.

The ROC of 0.7284 in Table 12 is acceptable by commercial standards. Other combinations of the application, macroeconomic and behavioural web related variables gave similar predictive accuracy. Table 13 presents the regression results for this model. Turning to the number of website visits, we can see that the PD is positively associated with the weights of evidence of both the median and maximum visits per month to the website. This result remains consistent with our previous results that larger weights of evidence for *Number of website visits* per month is associated with a higher PD. Overall, based on the ROC, coefficients and Odds, we conclude that models trained on these particular web variables are predictive. These results suggest that these models, when using website interactions specifically *Number of views of the customer's account, Number of visits to the company's website, Number of terms and conditions checked* and *Number of mobile devices used* are viable alternative models to predict credit risk when bureau and transactional data are not available.

This result is important for financial inclusion because it suggests that people with an open purchase account (whether used as a credit account or not and no matter how their application was evaluated in the first place) could be assessed by this type of model. With the availability of open banking, lenders might be willing to accept thin file customers by using merely application and very limited banking transaction variables and therefore, offer a small credit limit. That is, if they have been granted credit without either bureau data or transactional information related to the online retailer lender. These customers would subsequently have the possibility of building web behavioural data over a period of 12 months. As a result, lenders would be able to implement this type of model that will enable them to make decisions regarding increasing or decreasing the credit limit for thin files customers based on the new behavioural web related variables. These new covariates provide an alternative to models that include bureau and transactional variables in handling credit risk applications for additional credit (if they already have some) and potentially lenders can make responsible decisions by incorporating these variables.

---

[8] The assumptions are the samples come from normally distributed populations, randomness of the samples and equal variances of the populations. For a more complete description of the *t*-test, see [61; 62]. We use the Shapiro–Wilk normality test to confirm the normality assumption. Overall, the results shows that the *p*-values are larger than the significance level $\alpha = 0.01$ for all models across the different performance measures. We can conclude that the normality assumption holds for all models at $\alpha = 0.01$. The resulting *p*-values are available on request. The second assumption is met since the selected samples for training, validating and testing were randomly chosen from their underlying population. The third assumption, similarity of variances, was tested with the Levene's test. The results for the third assumption are similar to those observed for the first one. Thus, all the assumptions of the *t*-test were met.

[9] Although it may be that over-indebtedness precedes depressive symptoms rather than vice versa.

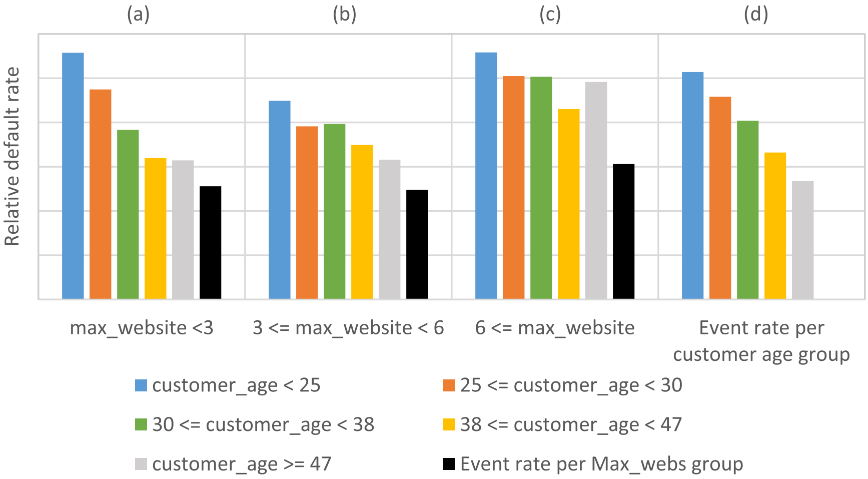

Fig. 3. Model with Lag12, A2. Relative default rate by maximum website visits per age group.

Table 12
Model with Lag 12 A5 without transactional and bureau variables.

		A5
	Application variables	2
	Macroeconomic variables	9
	Web behavioural variables	15
	Input variables	26
	Significant variables	13
KS	Train	0.3915
	Validation (in-time test)	0.3809
ROC	Train	0.7356
	Validation (in-time test)	0.7284

5. Conclusion

The aim of this paper is to show that the inclusion of new web browsing variables, such as *Number of website visits* to a retailer, the *Number of devices* used to access the lender's site, the *Number of account sessions* and *Number of terms and conditions views,* into survival analysis as predictors enhance the predictive accuracy of a PD model at account level.

Our results show first, that including various transformations of *Number of mobile devices* to make online purchases, *Number of visits* to an organisation's website, *Number of sessions* visiting the borrower's account, and *Number of terms and conditions views* into a survival scoring model enhances its predictive accuracy compared to one containing only conventional application, bureau and transactional variables. Although the addition of new web browsing variables as predictors of PD increases the area under the ROC curve by only 1.0%, their inclusion reduces relative misclassification costs of Type II error across a range of alternative cost scenarios. This indicates that the new behavioural web related variables that vary over time not only have predictive power but also provide promising information to reduce the costs of misclassification.

Second, our results highlight that the default rate for heavy website users is independent of age, which contrasts the well published negative

Table 13
Coefficient estimates from Model Lag 12 A5.

Category vars	Effect	Coefficient	Odds RE	Wald CS	*p-value*
Application	Brand_type_w	−1.054	0.349	449.550	<0.0001
	Customer_age_w	−0.753	0.471	219.570	<0.0001
Macro-economic	FTSE100_w	0.580	1.786	16.090	<0.0001
	GfK_w	0.631	1.879	50.180	<0.0001
	HousePI_w	−0.079	0.924	0.070	0.792
	IP_w	−0.179	0.836	5.740	0.017
	IUMB479_w	−0.089	0.915	0.820	0.365
	IUMCCTL_w	0.131	1.140	2.650	0.104
	LPMB4TF_w	−0.178	0.837	0.850	0.357
	RPI_w	0.394	1.483	1.440	0.230
	UER_w	0.072	1.074	0.860	0.354
Web behavioural	AverageperMonth_num_device_w	−1.940	0.144	56.700	<0.0001
	AverageperMonth_myaccount_w	0.630	1.878	4.430	0.035
	AverageperMonth_num_terms_w	9.899	0.000	3.940	0.047
	AverageperMonth_website_w	−0.430	0.651	2.320	0.128
	Max_num_devices_w	−1.581	0.206	33.450	<0.0001
	Max_num_myaccount_w	0.083	1.086	0.070	0.788
	Max_num_terms_w	−0.659	0.517	12.860	0.000
	Max_website_w	2.210	9.112	32.950	<0.0001
	Mean_num_device_w	−1.178	0.308	36.120	<0.0001
	Mean_num_myaccount_w	−0.544	0.580	5.170	0.023
	Mean_website_w	−1.631	0.196	43.040	<0.0001
	Median_num_devices_w	−0.369	0.692	1.790	0.181
	Median_num_myaccount_w	0.278	1.320	1.500	0.220
	Median_num_terms_w	−2.672	0.069	0.290	0.590
	Median_website_w	0.240	1.271	1.020	0.311

correlation between PD and age for application samples *as a whole*. We further find that heavy website users have a higher default probability than that for less frequent users. This finding is plausible given its consistency with psychology and financial well-being theories. Our results reveal that the new behavioural web variable *Maximum number of website visits* has a positive relationship with PD for individuals aged 30 years and over.

Third, we find that time-varying behavioural web related variables boost predictive accuracy in the long-term (over a 12 month horizon), but not in the short term (over a 3 month horizon). This result for a 12 month horizon provides a valuable tool for regulatory reporting under the *Regulatory use of system-wide estimations of PD, LGD and EAD* [72] and the *Guidance on credit risk and accounting for expected credit losses* issued by the Basel Committee on Banking Supervision [57] because the Basel Accord requires the estimation of PD in the course of one year. We also argue that this observation for a 12 month horizon is particularly useful because when the pattern is observed over a longer period, it is more likely to be indicative of a personality trait, whereas over a short term period a particular behavioural pattern may be merely temporary or circumstantial.

Fourth, we find models that include only predictors related to website interaction in place of transactional and bureau variables are highly predictive. Using this type of model could help a lender offer loans to some applicants who may not have a credit history.

Our findings are of interest to banks, retailers, lenders and in general the online retail credit industry because our results underscore that the incorporation of new web related behavioural variables (non-standard information) in credit risk models as predictors of the probability of a borrower defaulting increase accuracy in the PD predictions. Moreover, our results provide insights into other disciplines including financial well-being, personal financial risk management and online consumer behaviour. Given the increased generation of information on individuals, web related behavioural variables will be more important than ever in years to come.

Funding

This research was partially funded by a commercial organisation, which had no involvement in any phase of the preparation of this manuscript. We thank them for allowing us to use their data.

CRediT authorship contribution statement

Betty Johanna Garzon Rozo: Conceptualization, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jonathan Crook:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing. **Galina Andreeva:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Betty Johanna Garzon Rozo, Jonathan Crook, Galina Andreeva reports financial support was provided by Commercially confidential.

Data availability

The data that has been used is confidential.

References

[1] ONS, Economic trends in the retail sector, Great Britain: 1989 to 2021, Retrieved October 20, 2021 from, https://www.ons.gov.uk/economy/nationalaccounts/balanceofpayments/articles/economictrendsintheretailsectorgreatbritain/1989to2021#online-retail-in-the-uk-analysis-by-sector, 27 July 2021.

[2] S. Lessmann, B. Baesens, H.-V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, Eur. J. Oper. Res. 247 (1) (2015) 124–136.

[3] L. Thomas, J. Crook, D. Edelman, Credit scoring and its applications, Society for industrial and Applied Mathematics (2017).

[4] B. Klinger, L. Castro, P. Szenkman, A. Khwaja, Unlocking SME finance in Argentina with psychometrics, Inter-American Development Bank. Technical Note No. IDB-TN-532 (2013).

[5] B. Klinger, A. Khwaja, J. LaMonte, Improving credit risk analysis with psychometrics in Peru, Inter-American Development Bank. Technical Note No. IDB-TN-587 (2013).

[6] B. Klinger, A.I. Khwaja, C. Del Carpio, Enterprising Psychometrics and Poverty Reduction vol. 860, Springer, 2013.

[7] I. Arráiz, M. Bruhn, R. Stucchi, Psychometrics as a tool to improve credit information, World Bank Econ. Rev. 30 (Supplement_1) (2017) S67–S76.

[8] T.J. Dlugosch, B. Klinger, M. Frese, U.C. Klehe, Personality-based selection of entrepreneurial borrowers to reduce credit risk: two studies on prediction models in low-and high-stakes settings in developing countries, J. Organ. Behav. 39 (5) (2018) 612–628.

[9] C. Liberati, F. Camillo, Personal values and credit scoring: new insights in the financial prediction, J. Oper. Res. Soc. 69 (12) (2018) 1994–2005.

[10] V.B. Djeundje, J. Crook, R. Calabrese, M. Hamid, Enhancing credit scoring with alternative data, Expert Syst. Appl. 163 (2021), 113766.

[11] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, B. Baesens, APATE: a novel approach for automated credit card transaction fraud detection using network-based extensions, Decis. Support. Syst. 75 (2015) 38–48.

[12] S. De Cnudde, J. Moeyersoms, M. Stankova, E. Tobback, V. Javaly, D. Martens, Who cares about your Facebook friends, in: Credit Scoring for Microfinance (2015018), 2015.

[13] M. Lin, N.R. Prabhala, S. Viswanathan, Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending, Manag. Sci. 59 (1) (2013) 17–35.

[14] R. Ge, J. Feng, B. Gu, P. Zhang, Predicting and deterring default with social media information in peer-to-peer lending, J. Manag. Inf. Syst. 34 (2) (2017) 401–424.

[15] Q. Gao, M. Lin, R.W. Sias, Words Matter: The Role of Texts in Online Credit Markets, Journal of Financial and Quantitative Analysis (2018), https://doi.org/10.2139/ssrn.2446114 forthcoming, Available at SSRN: https://ssrn.com/abstract=2446114.

[16] O. Netzer, A. Lemaire, M. Herzenstein, When words sweat: identifying signals for loan default in the text of loan applications, J. Mark. Res. 56 (6) (2019) 960–980.

[17] R. Iyer, A.I. Khwaja, E.F. Luttmer, K. Shue, Screening peers softly: inferring the quality of small borrowers, Manag. Sci. 62 (6) (2016) 1554–1577.

[18] G. Dorfleitner, C. Priberny, S. Schuster, J. Stoiber, M. Weber, I. de Castro, J. Kammler, Description-text related soft information in peer-to-peer lending–evidence from two leading European platforms, J. Bank. Financ. 64 (2016) 169–187.

[19] T. Berg, V. Burg, A. Gombović, M. Puri, On the rise of fintechs: credit scoring using digital footprints, Rev. Financ. Stud. 33 (7) (2020) 2845–2897.

[20] J. San Pedro, D. Proserpio, N. Oliver, MobiScore: towards universal credit scoring from mobile phone data, in: International conference on user modeling, adaptation, and personalization, 2015.

[21] L. Ma, X. Zhao, Z. Zhou, Y. Liu, A new aspect on P2P online lending default prediction using meta-level phone usage data in China, Decis. Support. Syst. 111 (2018) 60–71.

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

[23] D. Björkegren, D. Grissen, Behavior revealed in mobile phone usage predicts credit repayment, World Bank Econ. Rev. 34 (3) (2020) 618–634.

[24] J. Zhou, C. Wang, F. Ren, G. Chen, Inferring multi-stage risk for online consumer credit services: an integrated scheme using data augmentation and model enhancement, Decis. Support. Syst. 113611 (2021).

[25] L. Gonzalez, Y.K. Loureiro, When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans, J. Behav. Exp. Financ. 2 (2014) 44–58.

[26] T. Lu, Y. Zhang, B. Li, Profit Vs. Equity? The Case of Cinancial Risk Assessment and a New Perspective on Alternative Data. https://ssrn.com/abstract=3758120, 2022.

[27] W. Wu, D. Xu, Y. Zhao, X. Liu, Do consumer internet behaviours provide incremental information to predict credit default risk? Econ. Polit. Stud. 8 (4) (2020) 482–499.

[28] A. Vissing-Jorgensen, Consumer credit: learning your customer's default risk from what (s)he buys, Available at, https://ssrn.com/abstract=2023238, 2021.

[29] W. Li, Y. Wu, Y. Zhang, T. Lu, Y. Xu, Y. Sun, Insights from niche markets: explainable and predictive value of consumption tendency in credit risk assessment, in: Available at SSRN 4136142, 2022.

[30] R. Kazemi, A. Mosleh, Improving default risk prediction using Bayesian model uncertainty techniques, Risk Anal. Int. J. 32 (11) (2012) 1888–1900.

[31] T. Bellotti, J. Crook, Credit scoring with macroeconomic variables using survival analysis, J. Oper. Res. Soc. 60 (12) (2009) 1699–1707.

[32] J.D. Kalbfleisch, R.L. Prentice, The Statistical Analysis of Failure Time Data vol. 360, John Wiley & Sons, 2011.

[33] D.W. Hosmer Jr., S. Lemeshow, Applied survival analysis: regression modelling of time to event data (1999), Eur. Orthodontic Soc. (1999) 561–562.

[34] J. Banasik, J.N. Crook, L.C. Thomas, Not if but when will borrowers default, J. Oper. Res. Soc. 50 (12) (1999) 1185–1190.

[35] M. Stepanova, L.C. Thomas, PHAB scores: proportional hazards analysis behavioural scores, J. Oper. Res. Soc. 52 (9) (2001) 1007–1016.
[36] M. Stepanova, L. Thomas, Survival analysis methods for personal loan data, Oper. Res. 50 (2) (2002) 277–289.
[37] T. Bellotti, J. Crook, Forecasting and stress testing credit card default using dynamic models, Int. J. Forecast. 29 (4) (2013) 563–574.
[38] V.B. Djeundje, J. Crook, Dynamic survival models with varying coefficients for credit risks, Eur. J. Oper. Res. 275 (1) (2019) 319–333.
[39] T. Bellotti, J. Crook, Loss given default models incorporating macroeconomic variables for credit cards, Int. J. Forecast. 28 (1) (2012) 171–182.
[40] V.B. Djeundje, J. Crook, Identifying hidden patterns in credit risk survival data using generalised additive models, Eur. J. Oper. Res. 277 (1) (2019) 366–376.
[41] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, M. Schmid, A review of spline function procedures in R, BMC Med. Res. Methodol. 19 (1) (2019) 1–16.
[42] S. Luo, X. Kong, T. Nie, Spline based survival model for credit risk modeling, Eur. J. Oper. Res. 253 (3) (2016) 869–879.
[43] A.C. Cameron, P.K. Trivedi, Microeconometrics: Methods and Applications, Cambridge University Press, 2005.
[44] S.P. Jenkins, Easy estimation methods for discrete-time duration models, Oxf. Bull. Econ. Stat. 57 (1) (1995) 129–138.
[45] G. Andreeva, S. Fine, Personality, financial knowledge and credit performance: the USA and the UK comparison, in: Presentation at the Credit Scoring & Credit Control XVI conference, Edinburgh, 2019. https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2019/07/CSCC-2019-psy-final.pdf.
[46] V.B. Djeundje, J. Crook, Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards, Eur. J. Oper. Res. 271 (2) (2018) 697–709.
[47] A. Đurović, Macroeconomic approach to point in time probability of default modeling–IFRS 9 challenges, J. Central Bank. Theory Pract. 8 (1) (2019) 209–223.
[48] Y. Li, Y. Li, Y. Li, What factors are influencing credit card customer’s default behavior in China? A study based on survival analysis, Phys. A Stat. Mech. Appl. 526 (2019), 120861.
[49] J.L. Breeden, J. Crook, Multihorizon discrete time survival models, J. Oper. Res. Soc. (2020) 1–14.
[50] J. Banasik, J. Crook, L. Thomas, Sample selection bias in credit scoring models, J. Oper. Res. Soc. 54 (8) (2003) 822–832.
[51] E.N. Tong, C. Mues, L.C. Thomas, Mixture cure models in credit scoring: if and when borrowers default, Eur. J. Oper. Res. 218 (1) (2012) 132–139.
[52] H. He, W. Zhang, S. Zhang, A novel ensemble method for credit scoring: adaption of different imbalance ratios, Expert Syst. Appl. 98 (2018) 105–117.
[53] J.P. Barddal, L. Loezer, F. Enembreck, R. Lanzuolo, Lessons learned from data stream classification applied to credit scoring, Expert Syst. Appl. 162 (2020), 113899.
[54] D.J. Hand, Good practice in retail credit scorecard assessment, J. Oper. Res. Soc. 56 (9) (2005) 1109–1117.
[55] R. Calabrese, G. Andreeva, J. Ansell, “Birds of a feather” fail together: exploring the nature of dependency in SME defaults, Risk Anal. 39 (1) (2019) 71–84.
[56] H.A. Abdou, J. Pointon, Credit scoring, statistical techniques and evaluation criteria: a review of the literature, Intel. Syst. Account. Finan. Manage. 18 (2–3) (2011) 59–88.
[57] BIS, Basel Committee on Banking Supervision: Consultative Document. Guidelines. Guidance on Credit Risk and Accounting for Expected Credit Losses, Retrieved October 15, 2021 from, https://www.bis.org/bcbs/publ/d350.pdf, 2015.
[58] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, CRC Press, 1994.
[59] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Ser. B Methodol. 36 (2) (1974) 111–133.
[60] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011.
[61] D.C. Howell, Statistical Methods for Psychology, PWS-Kent Publishing Co, 1992.
[62] T. Hill, P. Lewicki, P. Lewicki, Statistics: Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining, StatSoft, Inc, 2006.
[63] H. Odacı, Çikrikci, Ö., An exploration of the associations among internet use, depression, anxiety and stress among youths, Mediterran. J. Clin. Psychol. 5 (3) (2017).
[64] J.J. Block, Issues for DSM-V: internet addiction, American journal of Psychiatry 165 (3) (2008) 306–307.
[65] G.M. Cooney, J. Morris, Time to start taking an internet history? Br. J. Psychiatry 194 (2) (2009) 185.
[66] L.F. Dunn, I.A. Mirzaie, Consumer debt stress, changes in household debt, and the great recession, Econ. Inq. 54 (1) (2016) 201–214.
[67] D.A. Hojman, Á. Miranda, J. Ruiz-Tagle, Debt trajectories and mental health, Soc. Sci. Med. 167 (2016) 54–62.
[68] J. Crook, The demand and supply for household debt: A cross country comparison, in: The Economics of Consumer Credit, The MIT Press, Cambridge, 2006.
[69] J. Crook, Household debt demand and supply: a cross-country comparison, in: The Economics of Consumer Credit, 2006, pp. 63–92.
[70] A. Lusardi, O.S. Mitchell, V. Curto, Financial literacy among the young, J. Consum. Aff. 44 (2) (2010) 358–380.
[71] A. Demirguc-Kunt, L. Klapper, D. Singer, S. Ansar, J. Hess, The Global Findex Database: Measuring Financial Inclusion and the Fintech Revolution, World Bank Group, 2017.
[72] J.A.E. Flores, T.L. Basualdo, A.R.Q. Sordo, Regulatory Use of System-wide Estimations of PD, LGD and EAD: FSI Award 2010 Winning Paper, Financial Stability Inst., Bank for Internat. Settlements, 2010.

Betty Johanna Garzon Rozo is a Research Fellow of Predictive Analytics at the Credit Research Centre in the University of Edinburgh - Business School. She studied Industrial Engineering, holds a MSc in Industrial Engineering and received her PhD in Management Science, specialising in modelling Operational Risk in Banking, from the University of Edinburgh. Her research focuses on applications of advance statistical predictive and prescriptive analytics in banking risk. She has 14 years’ experience delivering powerful data-driven insight and analysis, which support decision making and generate value for global and local companies across several industries including online retail credit, banking, telecom and media. She has 6 years’ experience in consultancy services conducting research projects in credit risk modelling in the UK, which provides innovative and workable solutions to business problems.

Jonathan Crook is an Emeritus Professor of Business Economics at the University of Edinburgh - Business School. Jonathan studied Economics at Lancaster and Cardiff. He has been a Visiting Fulbright Postdoctoral Research Scholar at the McIntyre School of Commerce, University of Virginia, USA; a Visiting Fellow at the University of Warwick, UK, and a Visiting Fellow at the European University Institute, Florence. He is a Fellow of Financial Institutions Centre, Wharton School, University of Pennsylvania; an External Research Fellow of the Centre for Finance, Credit and Macroeconomics at the University of Nottingham, and has been elected a Fellow of the Royal Society of Edinburgh and made a Fellow of the Academy of Social Sciences.

He currently has a Leverhulme Emeritus Fellowship. He concentrates on two research areas: (i) Modelling of credit risk and operational risk. This research is motivated by modelling and risk management problems that lenders and fintechs face with a view to developing innovations that enable risk to be managed more effectively. (ii) The Economics of the consumer credit including the demand (consumption and finance models), the supply of credit and credit constraints using household level data. He has published over 100 scientific papers in top journals including *European Journal of Operational Research, Journal of Financial Services Marketing, Journal of the Royal Statistical Society: Series A, Review of Quantitative Finance and Accounting, Journal of the Operational Research Society, Journal of Financial Stability, Expert Systems with Applications.*

Galina Andreeva is a Professor in Management Science, head of Management Science and Business Economics Group, School Executive Member and Director of Credit Research Centre. She holds a PhD at Edinburgh University; MSc in Operational Research & Management Science at Edinburgh University and MA (Russia). Prior to her current position at the Associate Professor level, Galina worked at Bank of Scotland and held the prestigious ESRC post-doctoral fellowship. She was a Visiting Scholar at NYU Stern Business School and University Milano-Bicocca. Her research evolves around credit risk of individuals and small businesses (SMEs) using advanced statistical and machine-learning techniques.

She investigated profitability of individual accounts in consumer credit, explored the performance of UK SMEs through the credit crisis, inferred the effect of management capability on SME credit performance. She published in top journals - *Risk Analysis: An International Journal, European Journal of Operational Research, Journal of Financial Stability, Financial Accountability and Management, Expert Systems with Applications*. She led and collaborated on a number of consultancy projects in credit risk and risk analysis.