Credit Scoring with Alternative Data

Jonathan Crook[1], Raffaella Calabrese, Viani Djeundje, Mona Hamid
Credit Research Centre,
University of Edinburgh Business School

7 August 2019

1. Introduction

A substantial number of people in the world do not have an account with a financial institution. In 2017 Demirguc-Kunt et al. (2017) estimated that 1.7 billion adults (31% of the adult population) did not have an account with a financial institution nor a facility through a mobile money provider (Demirguc-Kunt, 2017). These adults are concentrated in developing countries, particularly in China (22.5m), India (190m) and Pakistan (100m). In many African countries the percentage without an account is estimated to be around 75%. The reasons for not having an account are varied including that a person does not wish an account or, if they do, they did not apply or, if they did apply, their application was declined. A recent survey found that of those surveyed 20% of those without an account said they did not have an account because they did not have adequate documentation. In the US 7% of adults were found not to have a financial or mobile financial account and in the UK it was 4%. However these data may not describe the proportion with credit since one can have an account without credit. The vast majority of financial institution lenders will only grant a loan to an applicant if the applicant has a credit score. In the US, Jennings (2015) using a FICO dataset estimated that 53m people could not gain a credit score because their credit records were insufficient or they did not have any records. Using the CFPB Consumer Credit Panel of 2010 and other sources, Brevoort et al (2016) put the figure at 45m with 9.9m having an insufficient credit history, 9.6m having a credit history that was too historic to be usable and 26m with no credit history. In many low income countries the reasons for not be able to gain a financial account are also due to lack of crucial characteristics necessary to gain a credit score.

Partly motivated by such high proportions of the adult population that cannot gain a score, a number of commercial organisations have developed scoring models that use non-traditional data. Examples include the use of rental data by Experian, and use of utility data, evictions, property values and other variables by FICO. But there is little detailed published analysis of the contributions of the components within these scores and they are applied typically in higher income countries. Other organizations, which have typically been start-ups, use different types of non-traditional data to estimate scoring models for application typically in lower income countries. Examples of the latter include Lenddo, Tala, Branch, among others. In the academic literature an increasing number of researchers have included non-conventional covariates into credit scoring models to assess their predictive power to distinguish between good and poor payers. This paper reports on experiments to assess the predictive accuracy of credit scoring models that use certain types of alternative data instead of, or as well as, conventional predictors.

In this paper we make two contributions. First, we show that using data on alternative characteristics, specifically characteristics of email useage and psychometrics, one can gain good separation between good payers and bad payers. Second, we show the relative contributions of these characteristics compared with demographic variables in a credit scoring model.

[1] Corresponding author

The next section reviews the empirical evidence on the use of alterative characteristics in credit scoring. Section three describes the data we used and sections four and five describe the analyses. In section five we comment on the implications of the results and the final section concludes.

2. Literature Review

Application credit scoring models predict whether a new applicant for a credit product will make the scheduled payments on time over a pre-defined outcome period that is usually 12 or 18 months. In traditional models the covariates (or inputs into a machine learning model) would include: items measured at the time of application such as years at address, years in employment, income, age , credit bureaux data such as repayment history on previous loans both at that institution and other institutions, proportion of the population in the postcode that default, etc. Behavioural scoring models are applied to accounts that have been open for a sufficient period of time for the analyst to assess characteristics of their use such as balance outstanding in the last 6 months and average expenditure on the account over the last 3 months. Applications and bureaux variables are also included. In both types of models the covariates may be described as socio-demographic and financial. If a model includes a variable relating to, for example, number of credit lines open in the last 3 months or whether an account has defaulted in the last 12 months, but there is no data for a new (or existing customer) for that variable a score cannot be obtained and in the case of a credit application it would be declined. Such variables are included in a very high proportion of scoring models. For example Jennings (2015) states that to gain a FICO score an individual must have at least one credit line open in the last 6 months. However the proportion of adults in certain countries, especially lower income countries, who have no credit history is relatively high.

Whilst not having had credit in the past may be due to previous credit risk assessments indicating too high a risk for a lender to grant a loan, this is not necessarily the case. For example people who migrate into a country, some new college students, people who do not use a financial account they already possess or in some cases people who have just never asked for a loan may also not have a sufficient credit history.

Since the late 2000s researchers have experimented with using covariates, other than conventional financial and socio demographic variables, to see if their inclusion, either instead of or as well as, conventional variables increases predictive accuracy or not. Variables relating to very different types of information have been used.

Several papers have used information contained in the textual description of the purpose a credit applicant would put a loan to. These studies commonly use data from a peer to peer loan applications. Dorfleitener et al (2016) considered the incidence of spelling errors, length of text and occurrence of types of keywords, but did not include any accuracy tests. Goa (2018) considered the readability, tone and occurrence of deception cues in loan purpose descriptions by borrowers on the Prosper peer to peer platform. Whilst no predictive accuracy statistics were included the authors found that a one standard deviation reduction in readability, a less positive tone or a higher level of deception cues were associated with an increase in the probability of default of up to 2.04%. Netzer et al (2018), who looked at both descriptions of purpose and of the credit applicant also on the Prosper platform, found that certain types of two word combinations and different types of word groupings were correlated with PD. They found an increase in area under the ROC curve (AUC) of 2.64 % when textual characteristics were added to financial and demographic characteristics and that the increase was greater for lower credit grades than for higher grades. Iyer at al. (2016) also use data from Prosper and compare the predictive accuracy of "soft" information in comparison with that from financial characteristics offered by borrowers. The "soft" information included whether the borrower posts a picture and the number of words used in the listing. They find that when including financial variables, including an Experian credit score, in the

model they gain an AUC of 0.710 and when they include the soft information as well the AUC increased to 0.714. Using data from an e-commerce furniture company in Germany Berg et al (2018) consider various indicators relating to a borrower's "digital footprint" such as whether the purchaser uses lower case when writing, errors when writing their email address as well as the operating system of the device used, and other variables. They find that adding digital footprint variables to a credit score alone increases the AUC from 0.680 to 0.728 for scorable customers and for unscorable customers they found that digital footprint variables gave similar predictive accuracy (0.683) to that gained by a bureau score for the scorable customers.

A number of papers have considered the predictive power of characteristics of mobile phone use. This is particularly important from a practical point of view because a much higher proportion of adults in lower income countries have and use a mobile phone than have a bank account. If information about mobile phone usage is predictive of default then it may be possible to use such data in place of financial and demographic variables for those for whom a credit score cannot be calculated.

Bjorkegren and Grissen (2018) use data from EFL relating to telecom loans made in a low income per head country. The characteristics of usage included measures of the periodicity of usage, the fraction of duration time spoken during a workday, variation in usage and the autocorrelation between calls and SMS messages. They found, perhaps surprisingly, that the phone indicators alone yielded a higher AUC than credit bureau data alone and that when phone indicators were added to bureau variables, the predictive accuracy increased, for example from 0.55 to 0.62.

Another strand of literature considers the predictive performance of psychometric variables. This is a poorly developed area and most of the empirical literature relates to loans to micro-entrepreneurs. In an early study Klinger et al (2013) used data relating to around 275 credit applicants from micro, small and medium sized enterprises in Peru. Sixty six psychometric variables were included (but not defined) and gave an AUC of 0.7 for a training sample. They also estimated a similar model for data from four African countries and tested it on the data from Peru and gained an AUC of 0.56 -0.58 for a default definition of 60 days or more. Unfortunately, testing a model estimated from entrepreneurs in a range of countries and suggesting that its accuracy can be assessed by using at test sample from another country is highly problematic. A later study by Arraiz et al (2017) used a larger sample from EFL and again un-identified psychometric variables to find that those who were accepted under a traditional credit scoring model and rejected on the psychometric model had a poorer repayment performance than those accepted on the traditional model. The sample consisted of banked entrepreneurs and the result did not apply to non-banked entrepreneurs. Dlogosch et al. (2017) used data relating to micro-entrepreneurs in Kenya in high stakes and low stakes situations. The psychometric variables included were interpreted as measures of conscientiousness, emotional stability, openness to experience and integrity. Unfortunately whilst an AUC of 0.67 was gained for a high stakes model the paper did not show the additional predictive power of including the psychometrics predictors. None of these papers show the increase in predictive performance when psychometric covariates are included as well as traditional financial variables.

In contrast, Liberati and Camillo (2018) extract six psychological constructs using principal components analysis from responses to a Semiometrie that had been administered by an Italian bank. The six dimensions were interpreted as being along the participation, duty/pleasure, attachment/detachment, sublimation/materialism, idealization/pragmatism and humility/sovereignty scales. Liberati and Camillo found that when these components are included in models that already included use of bank services, cash flow and a solvency score then the AUC increased considerably: from around 0.554 to around 0.850 (depending on the classifier used).

In summary alternative predictors in the form of characteristics of verbal descriptions from peer-to-peer sites and mobile phone usage have been found to have discriminatory power when classifying good and poor repayers. But the literature on psychometrics relates to micro entrepreneurs rather than consumers

and there are few papers that have estimated the predictive enhancement from using these types of variables in addition to others. In addition there is no paper that shows the predictive enhancement when characteristics of email activity by consumers are used and so none when they are separately and/or together with psychometric data. The aim of this paper is to evaluate the predictive performance of using psychometric variables and/or characteristics of email usage to predict the probability of default for consumers. We find that each type of predictor when used alone will yield a model with modest predictive accuracy, but when used together in an ensemble, both types of non-conventional variables enhance the predictive accuracy of demographic variables. We also find that the level of predictive accuracy when demographic and psychometric variables in particular are combined together in an ensemble model give predictive accuracy which is to a commercially acceptable level and so could, in principle, be used for credit applicanst for which no previous credit history is available.

3. Data

We use two groups of datasets which we refer to as "Set A" and "Set B". Both were supplied by Lenddo and originally sourced from a bank in Mexico and a bank in Nigeria, respectively. The data related to successful applications for micro credit where, for some of the cases, the repayment outcome was observed.

Set A comprises three datasets, one containing values for the demographic variables, one containing values for a selection of psychometric measures, and a third one containing values on other variables that we label "alternative data" and that relate to characteristics of email activity by the applicant. A summary of the size and structure of these datasets is shown in the upper part of Table 1 below. Notice that although the number of cases in the alternative dataset supplied seems larger than for the other two sets, the target variable (an indicator of default) was missing for the vast majority of them; only 442 cases had information on the target variable.

In each of these three datasets, there was a unique identifier (id) associated with each case. These ids were identical for the demographic and psychometric datasets. Regarding the alternative dataset, out of the 442 cases with a valid target value, the majority of cases (98%) were also found in the demographic dataset.

Set B comprises a single dataset of alternative data only; see the lower part of Table 1. This dataset is independent of those in Set A in the sense that it does not intersect (case wise) with any of the datasets from Set A. In addition, the overall default rate in three the datasets from Set A is much higher than that in Set B.

Table 1: Structure of the datasets received

		Demographic	Psychometric	Alternative data
Set A	number of variables	12	350	53
	number of cases	1826	1826	33091
Set B	number of variables	NA	NA	237
	number of cases	NA	NA	16358

4. Analysis of Set A

4.1 Data preparation

We first excluded cases for which the target variable was missing, separately for each dataset in Set A. Also, variables with negligible variance were filtered out, and underpopulated levels of categorical variables were merged into their closest level. The resulting datasets were used to estimate predictive models for the target variable. Descriptive statistics relating to the dataset A are given in the Appendix.

Two approaches were considered: in the first we estimated models using the observed data, whereas in the second approach we imputed values for missing data. Outputs from both approaches can be compared.

4.2 Analysis based on observed data

One possibility was to merge the three datasets based on the id field and then explore models on the combined dataset, excluding all cases involving missing records. However, an early investigation suggested that ensemble type models tend to perform better for these data. This is consistent with the literature (Lessmann et al 2015). Thus, a two stage procedure was adopted.

At the first stage, each dataset was considered separately and split randomly into a training (75%) and test (25%) set. Various model structures were then considered and models estimated using the training set. A variable was retained when it improved the overall model quality (as measure by the p-value or the Akaike Information Criterion). Logistic regression models built on appropriate subsets of variables and interactions were consistently among the best performing models in terms of simplicity and predictive power. Thus, at the end of this first stage, three logistic regression models were retained, one for each dataset. The estimated parameters in each model are shown in Tables 2, 3 and 4.

All of the values of the covariates are positive or zero so the marginal effects have the same sign as the coefficients of the variables in the logit, which are the values shown in Tables 2, 3 and 4. We comment first on the demographic variables. The only significant variables are having two dependents which reduces the probability of default, the number of hours worked per week which is associated with an increase in PD and gender with males having a lower PD. Apart from number of dependants these are not commonly used predictors in published papers. This partly for legislative reasons, for example lenders in western countries do not collect data on gender due to sex discrimination legislation. In the literature the number of dependents is correlated with the probability of default (Banasik & Crook 2007). The literature also suggests that older borrowers have a lower chance of default (for example, Djeundje and Crook 2019) but in this data age is not significant.

Now we consider the psychometric predictors. The two variables that record the applicant's preferences over funds immediately or in 3 months or in 6 month's time indicate the inter-temporal preferences of the applicant. There are at least three effects at work here. Receiving funds further into the future is less desirable because of their reduced purchasing power compared to today due to inflation. Second future receipts involve greater risk the funds may not be forthcoming. Thirdly the applicant might simply prefer funds now rather than in the future because he wishes to gain the utility from their use now rather than later. Our results suggest that the PD is greater for applicants who prefer funding now rather than in 3 months but not for those who prefer the funds now rather than in 6 month's time. There appears to be a non-linear relationship between the number of potential referees an applicant gives and PD. If he gives 3 the PD is lower but if he gives 5 it is higher. Perhaps with 5 the more risky applicant is trying to give the impression that he will be thought of as a good risk if he cites a large number of referees. The larger the number of people the applicant says steal in his community might be associated with the general degree of honesty in the community in which the applicant lives and appears positively correlated with higher default risk. Time taken to answer questions for which the applicant would be relatively sure of the answer might indicate a degree of gaming the answers with the longer time taken the more likely the respondent is working out the answer most likely to give a good credit score. The desire to have certain types of loans in 12 months appears to act as a deterrent to default. The greatest

effect of those considered, as measured by the coefficients on the dummy variables, indicating each preference, appears to be the desire to have a business loan followed by the desire to have a savings account, then a credit card and fourth home loan or mortgage. A business loan may be necessary for higher income whilst a savings account may indicate prudence and possibly saved income. The measure of moderation: a preference to spend unexpected income on the applicant's home or health rather than on entertainment may indicate a prudent attitude to expenditure whereas the median time taken to express a degree of agreement with certain statement may indicate someone who is more analytical and thoughtful.

Turning to the email characteristics, the greater the number of emails per year then higher the fraction of emails sent between midnight and 6:00am and the higher the fraction of emails sent or received from non-top financial product providers the greater the probability of default. On the other hand applicants with a greater the number of contacts or that send a higher the fraction of emails on Tuesdays, Thursdays, Saturdays and/or Sundays on average have a lower probability of default.

Table 2

Estimated coefficients for the submodel based on demographic data alone

Variable	*Coefficient*	*p-value*
Intercept	-1.6778	0.002
How long has had phone	0.0198	0.418
Number of dependents = 2 (coded 1, 0 otherwise)	-0.4387	0.010
Number of dependents = 6 (coded 1, 0 otherwise)	-0.978	0.606
Hours worked per week	0.0137	0.030
Work experience	-0.0117	0.689
Age in years	0.0079	0.555
Gender (male=1)	-1.9801	0.001
Income	-0.0001	0.287
*Age * gender*	0.0473	0.005
*How long has had phone * work experience*	-0.0036	0.099

Notes

The variables shown in Table 2 are those which were selected due to their contribution to the model. A variable is retained when it improves the overall model quality (as measure by the p-value or the Akaike Information Criterion).

Table 3

Estimated coefficients for the sub-model based on psychometric data alone

Variable	*Coefficient*	*p-value*
Intercept	-1.1474	0.005
Does the applicant have accounts at other banks or financial institutions (1=missing, 0 otherwise)	0.6864	0.026
Choice between a smaller amount of money now (coded 0) or a larger amount in 3 months (coded 1)	-0.3936	0.016
Choice between a smaller amount of money now (coded 0) or a larger amount in 6 months (coded 1)	-0.2338	0.154
How many persons may be contacted for a reference no=2 (coded 1, 0 otherwise)	-0.3612	0.064
no=3 (coded 1, 0 otherwise)	-0.5340	0.016
no=4 (coded 1, 0 otherwise)	-0.3459	0.169
no=5 (coded 1, 0 otherwise)	0.0909	0.700
How many people in your community steal from others?	0.0069	0.029
Time taken for applicant to answer simple questions such as birth date	0.0013	0.070
What products the applicant does not have but would like to gain in next 12 months:		
Credit card	-0.8884	0.004
Loan/overdraft	-0.6072	0.398
Home Loan/Mortgage	-0.7307	0.042
Vehicle Loan	-0.7942	0.017
Deposit accounts (current, saving or term)	-1.5107	0.002
Personal Loan	-1.0705	0.003
Business Loan	-1.5935	0.000
Other products	-0.5079	0.462
None	-1.1488	0.001
A test of whether applicant is a "team player" or an "individualist" (coded 1 if missing, 0 otherwise)	1.5878	0.054
A measure of moderation	0.0706	0.012
Median time taken to express level of agreement with a number of statements	0.0672	0.019
Similarity of answer to a repeated question	2.1017	0.040

Notes

The variables shown in Table 3 are those which were selected due to their contribution to the model. A variable is retained when it improves the overall model quality (as measure by the p-value or the Akaike Information Criterion).

Categories not retained in the model after selection procedure:

How many persons may be contacted for a reference: no=0 and no=1.

What products the applicant does not have but would like to gain: debit card, leasing.

Further details of questions asked

a) *How many persons may be contacted for a reference:* The applicant is asked "If more information is required for this application, who of the following could we contact? Please select all who may be contacted" Options are categorised by relationship to the applicant.
b) *How many people in your community steal from others:* responses on a scale 1 to 100.
c) *What products :* The variable records the first product the applicant mentions when asked this question.
d) *Team player:* The applicant is presented with two images and he is asked "Which blue person in the image is more like you?" The images are pulling a cart up a hill alone versus a person who is pulling a cart uphill with others.
e) *Measure of moderation:* Applicant has to allocate 10 coins from unexpected income to four categories: home, health, vacation or entertainment. The variable measures the ratio of number for home and health to number for vacation and entertainment.
f) *Median time taken to express level of agreement:* the possible levels of agreement are: "strongly agree", "agree", "neutral", "disagree", "strongly disagree". Example statement: "My life is mostly controlled by chance events".

Table 4

Estimated coefficients for the sub-model based on alternative data alone

Variable	*Coefficient*	*p-value*
Intercept	0.1687	0.705
Time in years to send last 2000 emails	0.6201	0.010
Number of contacts the applicant sent the last 2000 emails to	-0.0054	0.027
Average number of words the applicant used in the subject line of the last 2000 emails	-0.1434	0.108
Fraction of emails sent between 0000hrs and 0600hrs	1.7151	0.004
Fraction of emails sent between 1800hrs and 2400 hrs	1.4781	0.164
Fraction of emails that were sent on Tuesdays	-1.6544	0.048
Fraction of emails that were sent on Thursdays	-2.9411	0.006
Fraction of emails that were sent on Saturdays	-2.6813	0.015
Fraction of emails that were sent on Sundays	-3.6693	0.039
Fraction of emails that were sent to or received from non-top financial product providers	0.7980	0.087
Log of number of emails received from uber.com	23.8613	0.139
Log of number of emails received from uber	-24.0732	0.135

Notes

All fractions calculated over the most recent emails 2000 emails or however many were sent or received.

The variables shown in Table 4 are those which were selected due to their contribution to the model. A variable is retained when it improves the overall model quality (as measure by the p-value or the Akaike Information Criterion).

An illustration of the dimensions of these models and their predictive performance in terms of AUC is given in Table 4.

Table 5: Summary models from stage 1

	Demographic model	psychometric model	alternative model
Number of training cases	1370	1370	332
Number of parameters	11	23	13
AUC (training set)	63%	67%	67%
AUC (test set)	62%	60%	58%

At the second stage, aggregated logistic models were built by combining the scores from the models retained in Stage 1 (shown in Table 1). The parameters of these aggregated models were estimated based on a random sample (75%) of common cases in the three datasets, the other 25% were used to assess the predictive performance of the aggregated models. A summary of this performance is shown in Table 5. Overall, these aggregated models perform better than models from State 1.

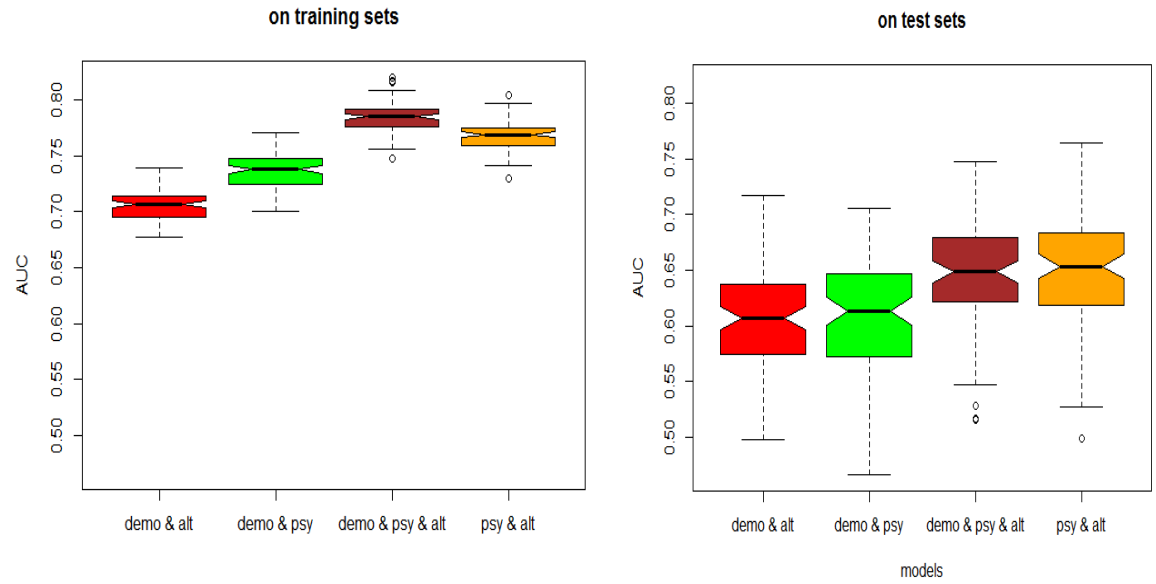
Table 5: Performance of the aggregated models from stage 2

	model (demographic +psychometric)	model (demographic +alternative)	model (psychometric +alternative)	model (demographic + psychometric + alternative)
AUC (training set)	66%	71%	71%	73%
AUC (test set)	75%	69%	67%	72%

During the analysis, it was found that the performance of these aggregated models tend to be sensitive to the training/test split. A simulation exercise was undertaken to investigate the magnitude of this sensitivity as follows. One hundred training/test sets were created by splitting at random the aggregated dataset. Each of the aggregated models shown in Table 5 was then fitted and assessed on these training/test sets. A comparative illustration of the outcome is shown in Figure 1. The length of the lines indicates the range of AUC values whilst the vertical dimension of a box indicates the interquartile range.

A number of conclusions can be drawn. First, these graphics confirm the sensitivity of the models with respect to the random train/test split, especially on the test sets. Second, the models show some signs of overfitting compared to the performance shown in Table 5. This is probably due to the fact that the structure of these models (i.e. selection of underlying variables and interaction terms) was not tailored to these individual training sets themselves, but instead was assumed to be the same structure as in Stage 1; see Table 1.

Figure 1: Sensitivity aggregated models with respect to the train/test split.



4.3 Analysis using imputed values

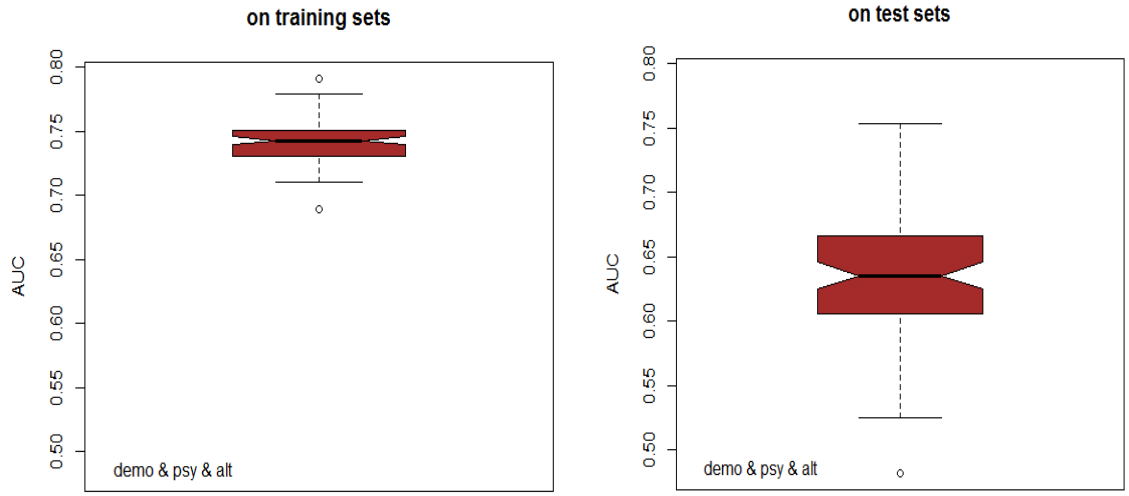
The second approach was to involve imputation of missing data in the modelling process. The starting point was to create a combined dataset by merging the three datasets from Set A using the id field (after removing rows with missing target value and filtering low variance variables in each dataset, separately). Note that this combined dataset contains a substantial number of missing data; first because each contributing dataset comes with its missing records, and second because the alternative variables were missing for a large proportion of cases in this combined dataset (indeed in the original alternative dataset, valid target value was available on 442 cases only).

In this second analysis, missing data were imputed. There are various imputation methods in the literature, from simple mean/mode substitution through to more advanced imputation methods. The method used in this analysis is the so called multiple imputation by chained equations (MICE) proposed by Raghunathon et al (2001). An attractive feature of this method is that it allows us to preserve not only the relations within the data but also the uncertainty about these relations. The method is as follows. Suppose we have a set of variables $(x_1, x_2, \ldots . x_p)$ and values are missing for some or all of them. Insert random values for those that are missing. Choose the variable with the fewest missing values, say it is x_1. Regress this on all of the other variables using only observed values of x_1, but observed and imputed values of all of the other variables. Predict the missing values of x_1. Then choose the variable with the next fewest missing values, say x_2, and regress the oberserved values of this variable on observed and imputed values of all of the other variables. Predict the missing values of x_2. Repeat this for all variables. Then repeat this 'cycle' a number of times. See Royston and White (2011).

Using this imputation method, 20 completed datasets were generated based on the underlying patterns and uncertainty in the original data. On each of these datasets, a logistic regression model was built using the demographic variables alone, and the 20 resulting models were averaged into one pooled demographic model following Little and Rubin (2002). Similarly, separate pooled psychometric and alternative models were constructed. The scores from these three pooled models were then ensembled together through a second layer logistic regression. An illustration of the performance of the resulting

model with respect to the random train/test split is shown in Figure 2. Overall, the performance is similar to the one without imputation described in Section a.

Figure 2: Prediction performance from the imputation based approach.



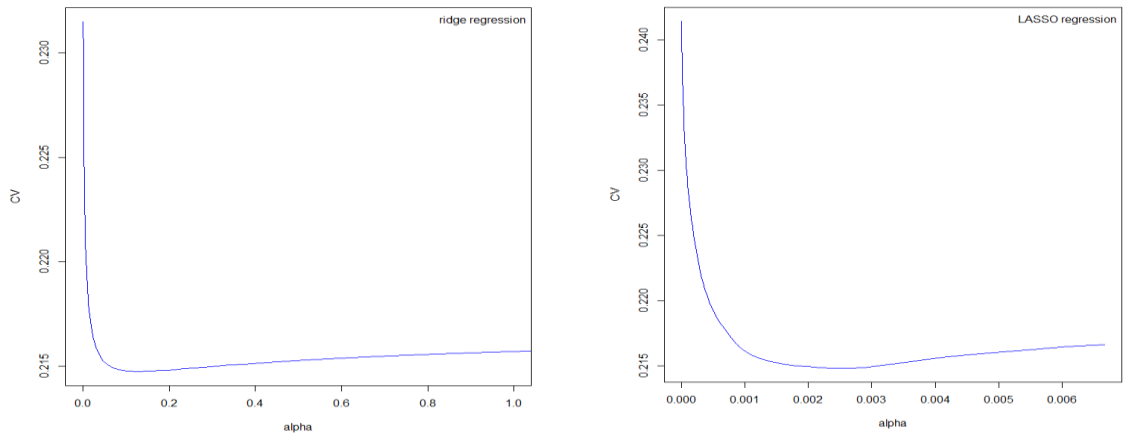
5. Analysis of Set B

Set B comprises a single dataset of 237 alternative variables. The dataset supplied had no missing records. However, the observed default rate in the dataset was very low (2%) relative to those in Set A. Given the large number of variables we have not presented summary statistics. We applied a wide range of classification methods, both statistical and machine learning. These included logistic regression, LASSO regression, ridge regression, extreme gradient boosting and deep neural networks. For each method, the dataset was randomly split into a training and a test set. The tuning of the underlying model parameters was based on the training set.

Ridge Regression is a form of penalised regression. It allows to prevent multicollinearity and reduce model complexity using regularisation techniques. Fitting a ridge regression model involves estimation of regression parameters and regularisation parameter (alpha). For a given value of alpha, the regression parameters can be estimated by maximization of the penalised likelihood. In this analysis, the optimal value of alpha was selected via cross validation (CV). The curve of the CV is shown on the left panel of Figure 3. The optimal value of alpha was 0.134. The final regression parameters were estimated based on this value of alpha.

LASSO regression is similar to ridge regression in the sense that complexity is simplified through regularisation. A graph of the CV as a function of the regularisation parameter is shown on the right hand side of Figure 3. But unlike ridge regression, the regularisation function used in LASSO regression tackles variable selection by shrinking the least important coefficients to zero. In this analysis for example, with the optimal regulation parameter of 0.0026, only 11 variables (out of 237) were retained.

Figure 3: Optimisation of regularisation parameter. Left: ridge regression. Right: LASSO



Extreme gradient Boosting is a machine learning technique comprising an ensemble of learners (usually decision trees) built in a hierarchical fashion. At each stage of the hierarchy, new trees are fitted to the residuals from previous stages and then used to update the ensemble. The performance of this method depends on the depth of the hierarchy, the scale of contribution of each tree, regularisation function, regularisation parameter, etc. Optimisation of these parameters was carried out using a combination of grid search and random initialisations. The performance of Extreme Gradient Boosting presented below in this section was achieved with a depth of 2, and contribution scale of 0.41 with no regularisation.

Deep Neural Network is a machine learning technique involving multiple layers between the input data and the output. The layers are made up of nodes and that is where computation takes place. In general, a node combines input from the data with a set of coefficients that either amplify or dampen the impact of that input. The performance of neural networks depends on a number of parameters, including the number of layers, the number of nodes within layers, learning rate, activation functions, etc. The prediction performance shown below in this section was obtained from a deep neural network with five layers (40-29-20-12-1); these were obtained by optimisation of the objective function using a combination of grid search and random initialisations.

For each method, the dataset was randomly split into a training and a test set. Models' parameters were tuned using the training set, and predictions were then assessed on the test set. A summary of the prediction performance of different classifiers is shown in Table 6. Due to the low fraction of default, attempt to oversample the default class was investigated. However, as can be seen from Table 6 oversampling yielded only minor improvements in the prediction performance.

We also extracted principal components from the covariates in the data and then fitted models to selected components. For each classifier, many scenarios involving different subsets of principal components were considered starting from the most important components. A comparative illustration of the importance of each principal component in terms of the percentage of variation explained is shown in Figure A1 in the Appendix. In this analysis, the subsets of principal components considered are those that were able to explain at least 60% of the variations in the original data.

Table 6: Performance of alternative data using different classification methods

	AUC train	AUC test
Logistic regression	56%	54%
LASSO	64%	57%
Ridge regression	67%	57%
Extreme Gradient Boosting	68%	62%
Oversampling1 + Extreme Gradient Boosting	82%	61%
Oversampling2 + Extreme Gradient Boosting	99%	62%
Neural Networks	97%	60%
PCA + Logistic regression	60%	54%
PCA + LASSO	60%	52%
PCA + Ridge Regression	60%	57%
PCA + Extreme Gradient Boosting	69%	62%
PCA + Neural Networks	73%	57%

Table 6 shows that the highest predictive accuracy was obtained from the Extreme Gradient Boosting, oversampling with Extreme Gradient Boosting and from the PCA with Extreme Gradient Boosting algorithms. All gave the same accuracy on the test set.

6. Discussion

Table 5 shows that the average predictive accuracies of demographic and psychometric variables in terms of AUC when using ensemble logistic regression was 0.75 and of demographic, psychometric and alternative data together it was 0.72. Both are somewhat higher than that derived from using either demographic variables alone (0.62 shown in Table 5), or psychometric variables alone (0.60 in Table 5), or alternative (characteristics of email usage) variables alone (0.58 in Table 5). In comparison, Berg et al (2018) gained an AUC of around 0.73 when using digital footprint characteristics, Iyer et al (2016) using whether a picture is submitted and text characteristics of peer to peer borrowers gained AUC values around 0.71. In psychometric studies Dlogosch (2017) gained an AUC of around 0.67 and Liberati and Camillo (2018) gained a figure of 0.85. In most cases we gain equal or higher predictive accuracy than published studies from psychometric or psychometric and alternative data. We could not find any papers that detail the predictive accuracy of using characteristics of email usage for predicting credit risk to compare with our results. However we must acknowledge some weaknesses in our work, in particular the small sample sizes. We hope to overcome this limitation in future work.

Turning to the implications of our findings, the use of alternative data, often mobile phone data and psychometrics is increasing in low income countries, especially for individuals who are otherwise unscorable. The predictive accuracy we have obtained suggests that these models, when using psychometric or email characteristic predictors are commercially viable as an alternative to models using financial data in the countries from which the data came. The practical implementation of models using these types of variables may however face challenges in Europe and the USA. For scorable applicants completing a psychometric profile as part of a credit application may be resisted due the time needed and the perceived invasiveness of the profile. There may also be concerns over the use of the information. In Europe the GDPR would require various permissions including that to use the data collected for model building. Unscorable applicants who would otherwise be rejected for credit may be much more willing to supply the necessary information. A further potential problem is that applicants

may learn to game psychometric profiles to gain a higher score. Mobile phone data is probably more difficult to game.

7. Concluding remarks

We conclude first that it is possible to use psychometric data and data relating to characteristics of email usage to increase the predictive accuracy of credit scoring systems. Second where access to standard credit scoring variables is difficult, the use of appropriate alternative data and psychometric characteristics of an applicant for a credit product can, on their own help a lender to score those who are credit invisible because sufficient data to enable a conventional credit score to be calculated is unavailable. Given the very large number of unscorable adults in the US (around 54 million) and in the African and Asian continents, these findings suggest a way of assessing the risk of lending to such large numbers of people which could potentially substantially increase the profits of lenders and increase demand in the economies where such loans could then be made.

References

Arraiz, I, Bruhn, M. and Stucchi, R. (2017) Psychometrics as a tool to improve credit information. The World Bank Economic Review, v30, Issue Supplemenet_1, S67-S76.

Banasik, J. & Crook, J. (2007) Reject inference, augmentation and sample selection. European Journal of Operational Research, 183, 1582-1594.

Berg, T., Burg, V., Gombovic, A., and Puri, M. (2018) *On the rise of the FinTechs – credit scoring using digital footprints*. Federal Deposit Insurance Corporation, Center for Financial research, Working Paper 2018-04.

Bjorkegren, D. and Grissen, D. (2018) *Behaviour revealed in mobile phone usage predicts loan repayment*. Department of Economics, Brown University Working Paper, SSRN 2611775

Brevoort, K.P. Grimm, P., and Kambara, M. (2016) Credit invisibles and the unscored. *Cityscape*, 18(2), 9-34.

Demirguc-Kunt, A, Klapper, L., Singer,D., Ansar, S. & Hess, J (2017) *The Global Findex Database* 2017.

Djeundje, V. and Crook, J. (2019) Identifying hidden patterns in credit risk survival data using Generalised Additive Models. *European Jo of Operational Research, 277(1), 366-376.*.

Dlogosch, T.J., Klinger, B. and Frese, M. (2017), Personality-based selection of entrepreneurial borrowers to reduce credit risk: two studies on prediction models in low- and high-stakes settings in developing countries. *Journal of Organisational Behaviour*, 39, 612-628.

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

Goa, Q., Lin, M. and Sias, R. (2018) *Words matter: the role of texts in online credit markets*. Available at SSRN 2446114.

Iyer, R., Khwaja, A.I., Luttmer, E.F.P. (2016) Screening borrowers softly: inferring the quality of small borrowers. *Management Science*, 62(6) 1554-1577.

Jennings, A, (2015) *Expanding the credit eligible population in the USA: a case study*. Presentation at the Credit Scoring and Credit Control XIV conference, Edinburgh.

Klinger,B., Khwaja, A.I., and LaMonte, J. (2013) *Improving credit risk analysis with psychometrics in Peru*. Inter-American Development Banks, Technical Note No IDB-TN-587.

Lessmann, S. ,Baesens, B., Seow, H-V. and Thomas, L.C. (2015) "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136.

Liberati, C. and Camillo, F. (2018) Personal values and credit scoring: new insights in the financial prediction. *Journal of the Operational Research Society*, 69(12), 1994-2005.

Little, R.J.A. and Rubin D. B. (2002) *Statistical analysis with missing data*. New York: Wiley.

Netzer, O., Lemaire,A. and Herzenstein, M. (2019) *When words sweat: identifying signals for loan default in the text of loan applications*. Columbia Business School Research Paper 16-83, available at SSRN 2865327..

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1), 85-95.

Royston, P. and White, I.R. (2011) Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of Statistical Software*, 45(4), 1-20.

Appendix

Further details of the analysis of the datasets in Set A:

Table A1

Mean values of predictors

Variable name	**Mean**	**# valid cases**
Socio-demographic		
How long phone	11.28	1826
Number of dependents	1.058	1826
Weekly workhours slide	44.98	1812
Workexperience slide	9.98	1823
Age (years)	33.74	1826
Gender	0.499	1826
Income_cns_dol	987.25	1802
Psychometric		
Has accounts at other financial institutions	1.3370	1751
Money now or in three months	1.5991	1826
Money now or in six months	1.6358	1826
Number of contacts	2.5290	1826
Time taken to answer simple questions	111.57	1826
Financial products desired but not yet have	17.97	1826
Team player or individualist	0.8471	1818
Measure of moderation	3.0253	1826
Median time to express agreement	7.0175	1826
Similarity of answer to repeated question	0.0069	1785
Variable name	**Mean**	**# valid cases**
Alternative data		
Time in years to send last 2000 emails	0.7306	442
Number of contacts the applicant sent the last 2000 emails to	40.64	442
Average number of words the applicant used in the subject line of the last 2000 emails	3.8770	367
Fraction of emails sent between 0000hrs and 0600hrs	0.4006	442
Fraction of emails sent between 1800hrs and 2400 hrs	0.1113	442
Fraction of emails that were sent on Tuesdays	0.1567	442
Fraction of emails that were sent on Thursdays	0.1504	442
Fraction of emails that were sent on Saturdays	0.1103	442
Fraction of emails that were sent on Sundays	0.0524	442
Fraction of emails that were sent to or received from non-top financial product providers	0.4790	367
Log of number of emails received from uber.com	1.1709	442
Log of number of emails received from uber	1.1758	442

Figure A1: Relative importance of the principal components (Set B).