Enhancing credit scoring with alternative data

Enhancing Credit Scoring with Alternative Data

Viani B. Djeundje, Jonathan Crook, Raffaella Calabrese, Mona Hamid
Credit Research Centre, University of Edinburgh

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

viani.djeundje@ed.ac.uk, j.crook@ed.ac.uk, raffaella.calabrese@ed.ac.uk,
mona.h.66@hotmail.com

Corresponding author:
j.crook@ed.ac.uk
Tel: +44 (0)131 650 3802

Enhancing Credit Scoring with Alternative Data

Viani B. Djeundje, Jonathan Crook, Raffaella Calabrese, Mona Hamid

Abstract

Hundreds of millions of people in low-income economies do not have a credit or bank account because they have insufficient credit history for a credit score to be ascribed to them. In this paper, we evaluate the predictive accuracy of models using alternative data, that may be used instead of credit history, to predict the credit risk of a new account. Without alternative data, the type of data that is typically available is demographic data. We show that a model that contains email usage and psychometric variables, as well as demographic variables, can give greater predictive accuracy than a model that uses demographic data only and that the predictive accuracy is sufficiently high for the demographic and email data to be used when conventional credit history data is unavailable. The same applies if merely psychometric data is included together with demographic data. However, we show that different randomly selected training: test sample splits give a wide range of predictive accuracies. In the second part of the paper, using two datasets that include only email usage as a predictor, we compare the predictive performances of a wide range of machine learning and statistical classifiers. We find that some classifiers applied to these alternative predictors give sufficiently accurate predictions for these variables to be used when no other data is available.

Keywords

Credit Scoring; Alternative Data; Banking Risk.

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. These adults are usually concentrated in developing countries, particularly in China (204m), India (357m) and Indonesia (102m). 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. Demirguc-Kunt found that of those surveyed 20% of those without an account said they did not have an account partly 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 being 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. However, 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 organisations which have typically been start-ups, use different types of non-traditional data to estimate application scoring models typically in lower-income countries. Examples of the latter include Lenddo, Tala and Branch, among others. In the academic literature, an increasing number of researchers have included non-conventional covariates into credit scoring

We would like to thank the ESRC for funding this work through the University of Edinburgh Impact Accelerator Account.

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.

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. Based on our knowledge, 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 used together with psychometric data. Furthermore, whilst psychometric variables have been used in the literature primarily in scoring models for small businesses, they have not been used for consumers.

In this paper, we make two main contributions. First, we show that by using data on alternative characteristics, specifically features of email usage 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.

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 in an ensemble model give predictive accuracy which is to a commercially acceptable level and so could, in principle, be used for credit applicants for which no previous credit history is available.

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

1 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 and credit bureaux data such as repayment history on previous loans both at that institution and other institutions, the proportion of the population in the postcode that default, etc. (Thomas et al., 2017). Behavioural scoring models are applied to accounts that have been open for a sufficient period for the analyst to assess characteristics of their use such as balance outstanding in the last six months and average expenditure on the account over the last three months (Thomas, 2000). Application and bureaux variables are also included (Hand and Henley, 1997). In both types of models, the covariates may be described as socio-demographic and financial (Meier and Sprenger, 2010). Whilst some covariates may include a missing value category, for others a missing value may result in an application being rejected. For example, if a model includes a variable relating to, for example, a number of credit lines open in the last three 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 very often would be declined. Applicants with such missing values are sometimes described as having 'no file' or a 'thin file'. 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 six months. Agarwal et al. (2019), Carol and Rehmani (2017), Brevoort et al. (2015), San Pedro et al. (2015) and Scheider and Schutt (2007) make a similar point. This is particularly common in lower-income countries where the proportion of adults 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 (Makela et al. 1993), 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 (De Cnudde et al. 2019, Oskarsdottr et al., 2019 among others) 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.

In this paper, we concentrate on psychometric variables and variables relating to email usage. The literature that considers the predictive performance of psychometric variables is limited, and much of the empirical literature relates to loans to micro-entrepreneurs.

In an early study, Meier and Sprenger (2007) using a laboratory experiment, found that impatience was correlated with default. 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 loans to entrepreneurs in a range of countries and suggesting that its accuracy can be assessed by using a 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 shows the increase in predictive performance when psychometric covariates are included as well as traditional financial variables.

In contrast, Liberati and Camillo (2018) extracted 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 with the participation, duty/pleasure, attachment/detachment, sublimation/materialism, idealisation/pragmatism and humility/sovereignty scales. Liberati and Camillo found that when these components are included in models that already included the 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. However, 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.

We cannot find any papers that have related the probability of default to characteristics of email usage. There are however papers that relate to aspects of the use of mobile phones as well as studies relating to characteristics of text that is used to describe the use of a loan and papers that consider Facebook data and social network information. These are summarised in Table A2 in the Appendix.

2 Data

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

		Demographic	**Psychometric**	**Alternative**
Ensemble A	*Number of variables*	12	350	53
	Number of accounts (rows)	1,826	1,826	33,091
Ensemble B	*Number of variables*	NA	NA	237
	Number of accounts (rows)	NA	NA	16,358

Table 1: Structure of the Lenddo datasets. There are three datasets in Ensemble A, but only one dataset in Ensemble B. The dataset in Ensemble B has no relationship with the datasets in Ensemble A.

Ensemble A comprises three datasets as follows. The first dataset consists of information on 12 demographic variables, the second consists of information on 350 psychometric variables, and the third dataset consists of information on 53 alternative variables labelled as "alternative data". In addition, each of these three datasets contains a repayment outcome taking value one if the account holder was unable to repay the loan, and zero otherwise. This repayment outcome is our target variable.

The alternative data consists of features of the customers' email activity. A summary of the size and structure of these three datasets in Ensemble A is shown in the upper part of Table 1. We make two remarks.

First, although the number of accounts in the alternative dataset supplied seems larger than for the other two datasets, the value of the target variable was missing for the vast majority of them; only 442 accounts had a non-missing repayment outcome. Second, account-wise, these three datasets in Ensemble A are not mutually exclusive. In

particular, the accounts in the demographic and psychometric datasets are the same; but regarding the alternative dataset, out of the 442 accounts with a non-missing repayment outcome, the vast majority of cases (98%) were also found in the demographic dataset.

We turn to Ensemble B. This comprises a single dataset of alternative variables; see the lower part of Table 1. The construction of the alternative variables in Ensemble B is different from that of those in Ensemble A. In addition to the alternative variables, this dataset contains a target variable representing the indicator of default.

Account-wise, this unique dataset in Ensemble B does not intersect with Ensemble A in the sense that none of the accounts in Ensemble B was found in any of the datasets from Ensemble A. In addition, the overall default rate (18%) in the three datasets from Set A is much higher than that in Ensemble B (2%).

3 Boosting credit scoring with alternative data

Demographic variables play an important role in credit scoring. The main objective of this section is to explore and quantify the predictive improvement if any, that alternative data can add to standard scoring models built on more traditional data. This will be achieved using the three datasets from Ensemble A, introduced in Section 2.

These datasets, as provided by Lenddo, required extensive cleaning. We started by excluding cases for which the target variable was missing, separately for each of the three datasets. Also, variables with negligible variance were filtered out, and underpopulated levels of categorical variables were merged into a neighbouring category. The resulting datasets were used to estimate predictive models for the target variable. Descriptive statistics relating to the three datasets in Ensemble A are given in Table A1 in the Appendix.

Two approaches were considered to analyse these data. In the first, we estimated models using the observed data, and we present the main results in this section. In the second approach, we imputed values for missing data; the results are shown in Appendix C. The outputs from both approaches can be compared. The analysis was carried out in `R`.

One possibility was to merge the three variable sets based on the id field and then explore models on the combined set, excluding all cases with 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.

3.1 Stage 1: Benchmark models for demographic, psychometric and alternative data

At the first stage, each set 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 from the final logistic models fitted separately to each dataset in Ensemble A are shown in Tables 2, 3 and 4. As all the values of the covariates are positive or zero, the marginal effects have the same sign as the variable coefficients in the logit model, which are the values shown in Tables 2, 3 and 4.

As can be seen from Table 2 there are only four demographic variables that are significant at 5%: number of working hours per week, gender, and number of dependents and the interaction of age and gender. As expected, the first variable shows a positive association with the probability of default (PD). For the second one, males have a lower PD. Apart from a number of dependants, these are not commonly used predictors in published papers. This is partly for legislative reasons, for example, lenders in western countries do not collect data on gender due to gender discrimination legislation. However, they can be used in some countries outside of Europe and the US. In the literature, the number of dependants is correlated with the probability of default (Banasik and Crook 2007, Tong et al. 2012). 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. Literature also suggests that additional work experience (Tong et al. 2012) and income reduce PD, and we find that in our data too, although neither is significant. Lack of significance may reflect collinearity, but we are interested in predictive accuracy, and so we are not so concerned about collinearity.

Now we consider the psychometric predictors. The two variables that record the applicant's preferences over funds immediately rather than in three months or in six months' time indicate the inter-temporal preferences of the applicant. There are at least three

Table 2: Estimated coefficients for the submodel based on only demographic variables.

	Coefficient	sdt. error	p-value
Intercept	-1.6778	0.5503	0.0023
How long has had phone	0.0198	0.0245	0.4183
Number of dependents = 2 (coded 1, 0 otherwise)	-0.4387	0.1699	0.0098
Number of dependents = 6 (coded 1, 0 otherwise)	-0.0978	0.1898	0.6063
Hours worked per week	0.0137	0.0063	0.0304
Work experience	-0.0117	0.0293	0.6887
Age in years	0.0079	0.0134	0.5550
Gender (male=1)	-1.9801	0.5945	0.0009
Income	-0.0001	0.0001	0.2870
*Age * gender*	0.0473	0.0168	0.0049
*How long has had phone * work experience*	-0.0036	0.0022	0.0986

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).

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. Secondly, future receipts involve a greater risk the funds may not be forthcoming. Thirdly the applicant might simply prefer funds now rather than in the future because he/she 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 three months but not for those who prefer the funds now rather than in six months. There appears to be a non-linear relationship between the number of potential referees an applicant gives and PD. If he/she gives three, the PD is lower but if he/she gives more than three PD is unaffected. Perhaps with more than three, the more risky applicant is trying to give the impression that he/she will be thought of as a good risk if he/she cites a large number of referees.

The larger the number of people the applicant says steal in his/her 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 the 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.

Table 3: Estimated coefficients for the sub-model based on psychometric data alone.

	coefficient	sdt error	p-value
Intercept	-1.1474	0.4078	0.0049
Does the applicant have accounts at other banks or financial institutions	0.6864	0.3082	0.0260
Choice between a smaller amount of money now (coded 0) or a larger amount in 3 months (coded 1)	-0.3936	0.1630	0.0157
Choice between a smaller amount of money now (coded 0) or a larger amount in 6 months (coded 1)	-0.2338	0.1642	0.1544
How many persons may be contacted for a reference:			
no=2 (coded 1, 0 otherwise)	-0.3612	0.1948	0.0638
no=3 (coded 1, 0 otherwise)	-0.5340	0.2206	0.0155
no=4 (coded 1, 0 otherwise)	-0.3459	0.2512	0.1686
no=5 (coded 1, 0 otherwise)	0.0909	0.2363	0.7004
How many people in your community steal from others?	0.0069	0.0032	0.0290
Time taken for applicant to answer simple questions such as birth date	0.0013	0.0007	0.0700
What products the applicant does not have but would like to gain in next 12 months:			
Credit card	-0.8884	0.3101	0.0042
Loan/overdraft	-0.6072	0.7190	0.3983
Home Loan/Mortgage	-0.7307	0.3585	0.0415
Vehicle Loan	-0.7942	0.3311	0.0165
Deposit accounts (current, saving or term)	-1.5107	0.4791	0.0016
Personal Loan	-1.0705	0.3534	0.0025
Business Loan	-1.5935	0.4176	0.0001
Other products	-0.5079	0.6909	0.4623
None	-1.1488	0.3476	0.0010
A test of whether applicant is a "team player" or an "individualist" (coded 1 if missing, 0 otherwise)	1.5878	0.8249	0.0543
A measure of moderation	0.0706	0.0281	0.0120
Median time taken to express level of agreement with a number of statements	0.0672	0.0286	0.0189
Similarity of answer to a repeated question	2.1017	1.0243	0.0402

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).

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".

The greatest marginal 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 a 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 a certain statement may indicate someone who is more analytical and thoughtful.

Table 4: Estimated coefficients for the submodel based on only alternative data.

	Coefficient	sdt. error	p-value
Intercept	0.1687	0.4448	0.7046
Time in years to send last 2000 emails	0.6201	0.2408	0.0100
Number of contacts the applicant sent the last 2000 emails to	-0.0054	0.0025	0.0274
Average number of words the applicant used in the subject line of the last 2000 emails	-0.1434	0.0893	0.1082
Fraction of emails sent between 0000hrs and 0600hrs	1.7151	0.5920	0.0038
Fraction of emails sent between 1800hrs and 2400 hrs	1.4781	1.0625	0.1642
Fraction of emails that were sent on Tuesdays	-1.6544	0.8356	0.0477
Fraction of emails that were sent on Thursdays	-2.9411	1.0595	0.0055
Fraction of emails that were sent on Saturdays	-2.6813	1.0981	0.0146
Fraction of emails that were sent on Sundays	-3.6693	1.7810	0.0394
Fraction of emails that were sent to or received from non-top financial product providers	0.7980	0.4661	0.0869
Log of number of emails received from uber.com	23.8613	16.1325	0.1391
Log of number of emails received from uber	-24.0732	16.1243	0.1354

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).

Turning to the email characteristics, on the one hand, Table 4 shows that the probability of default is positively associated with the fraction of emails sent between midnight and 6:00 am, as well as the fraction of emails sent or received from non-top financial product providers. On the other hand applicants with a greater number of contacts or that send a higher fraction of emails on Tuesdays, Thursdays, Saturdays and/or Sundays on average have a lower probability of default, as do those who send a greater number of emails per year. The predictive performance of each model is shown in Table 5 where the choice of covariates was made to optimise predictive performance in each case

Table 5 shows that in terms of the probability that a classifier will give a higher PD to a randomly selected default than to a randomly selected non-default (AUC), the model containing only demographic variables gives a better performance than that containing psychometric variables and that containing alternative variables, whereas the Pseudo-R^2 suggests that the alternative variables model has a greater fit. The Hosmer-Lemeshow

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
Pseudo-R^{2*}	3.32%	6.92%	9.18%
H-L test (p-value)†	0.4849	0.5457	0.1385
AUC (training set)	0.6311	0.6775	0.6745
AUC (test set)	0.6238	0.6007	0.5823

test shows that we cannot reject the null hypothesis that the observed event rates equal the expected ones.

3.2 Stage 2: An ensemble model for demographic, psychometric and alternative data

At the second stage, aggregated logistic models were built by combining the scores from the models retained in Stage 1 (shown in Table 5). The parameters of these aggregated models were estimated based on a random sample (75%) of common cases in the three datasets, and the other 25% was used to assess the predictive performance of the aggregated models. A summary of this performance is shown in Table 6. Overall, based on the AUC, these aggregated models perform better than models from Stage 1.

Table 6: Performance of the aggregated models from stage 2.

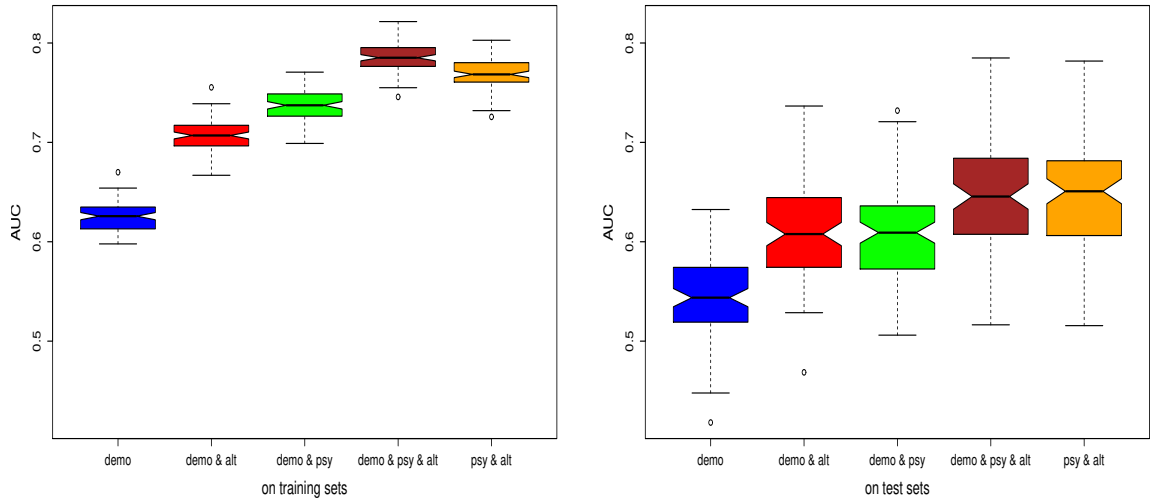
	Model (demographic+ psychometric)	Model (demographic+ alternative)	Model (psychometric+ alternative)	Model (demographic+ psychometric+ alternative)
AUC (training set)	0.6574	0.7116	0.7198	0.7253
AUC (test set)	0.7514	0.6910	0.6727	0.7212

During the analysis, it was found that the performance of these aggregated models tended 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

*McFadden's pseudo-R squared (Mittlbock and Shemper, 1996)

†Hosmer-Lemeshow goodness of fit test, the null hypothesis being that the model fits the data well.

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



models shown in Table 6 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 6. 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.

4 Credit scoring in the absence of traditional data

In the previous sections, we looked at how alternative data can help to improve the predictive performance of standard credit scoring models built on more traditional data such as demographic data. In practice, however, there are situations where access to traditional data is proven difficult. For example, in Ensemble B introduced in Section 2, demographic data was not available. In such situations, is the alternative data enough on its own to predict defaults?

To answer this question, we consider the alternative dataset from Ensemble B, and the alternative dataset from Ensemble A, separately. For each of them, a number of machine learning methods are implemented and used to predict defaults.

For each method, the relevant dataset was randomly split into a training (75%) and a test set (25%). The estimation of the underlying parameters of the models was carried out using the training set. In some instances, the training data was further split into two parts, in which case the first part was used to estimate models parameters and the second part used to tune hyper-parameters.

4.1 Predicting loan defaults using the alternative data from Ensemble B

We start with the alternative dataset from Ensemble B. This dataset contains observations of 237 alternative variables on 16,358 credit accounts. There are no missing records in this dataset. However, the observed default rate in the dataset was very low (2%) relative to that in Ensemble A (18%). 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. This included logistic regression, ridge regression, LASSO regression, extreme gradient boosting and deep neural networks. A description of these methodologies is provided in Appendix B.

A summary of the prediction performance of different classifiers is shown in Table 7. Due to the low fraction of defaults, we experimented with oversampling the defaulted cases. However, as can be seen from Table 7, oversampling yielded only minor improvements in prediction performance.

We also extracted principal components from the empirical covariance matrix of the covariates and then fitted classifiers to selected principal 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 A4 in the Appendix. In this analysis, the subsets of principal components retained are those that were able to cumulatively explain at least 60% of the variations in the original data.

4.2 Predicting loan defaults using the alternative data from Ensemble A

We turn to the alternative data from Ensemble A. Due to the small size of this dataset, only a reduced number of machine learning methods were investigated. The predictive performance of these methods is presented in Table 8.

[‡]In addition to the machine learning methods shown in Table 7, other algorithms such as decision trees and random forests were considered. But their prediction performance of this dataset was found to be close to random guess.

Table 7: Predictive performance of alternative data from Ensemble B using different classification methods[‡]

Method	AUC train	AUC test
Logistic regression	0.5605	0.5384
LASSO	0.6364	0.5651
Ridge regression	0.6672	0.5656
Extreme Gradient Boosting	0.6831	0.6203
Oversampling1 + Extreme Gradient Boosting	0.8226	0.6108
Oversampling2 + Extreme Gradient Boosting	0.9912	0.6251
Neural Networks	0.9776	0.5956
PCA + Logistic regression	0.6034	0.5368
PCA + LASSO	0.6032	0.5196
PCA + Ridge regression	0.6032	0.5745
PCA + Extreme Gradient Boosting	0.6938	0.6241
PCA + Neural Networks	0.7308	0.5748

The results in this table show that penalised-type regression (LASSO) turns out to perform quite well for this dataset compared to the standard logistic model. In addition, this table also shows that models built on appropriate subsets of the principal components tend to be better overall than models build on raw variables in the dataset; this remark applies to both the standard logistic form as well as the penalised forms.

Note that in this analysis of the alternative data from Ensemble A shown in Table 8, this logistic model is the same that which was used as part of the ensemble model presented in Section 3.2. Thus, the result in Table 8 signals that the ensemble model built in Section 3.2 could potentially be improved, for example by using the LASSO model build on principal components in place of the simple logistic model.

5 Discussion

Table 6 shows that the predictive accuracies of demographic and psychometric variables in terms of AUC when using ensemble logistic regression was 0.7514 and of demographic, psychometric and alternative data together it was 0.7212. Both are somewhat higher

[§]Only a reduced number of machine learning methods were fitted to the alternative dataset from Ensemble A due to the small size of this dataset.

Table 8: Predictive performance of alternative data from Ensemble A using different classification methods[§]

Method	AUC train	AUC test
Logistic regression	0.6745	0.5823
LASSO	0.6298	0.6217
Ridge regression	0.6204	0.6037
Extreme Gradient Boosting	0.6084	0.5984
PCA + Logistic regression	0.6231	0.6318
PCA + LASSO	0.6766	0.6318
PCA + Ridge regression	0.6749	0.6402
PCA + Extreme Gradient Boosting	0.6027	0.6067

than that derived from using either demographic variables alone (0.6238 shown in Table 5), or psychometric variables alone (0.6007 in Table 5), or alternative (characteristics of email usage) variables alone (0.5823 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. Studies utilising mobile phone records gain higher predictive accuracy. For example, Oskarsdottir et al. (2019) gained an AUC of 0.92 and Tan et al. (2015) gained 0.76. But this is not the case for other studies. For example Agarwal et al. (2019) gain an AUC of only 0.49 when using logistic regression. In some respects, it is difficult to compare AUC results across studies because they all contain different additional variables that often contribute significantly to predictive accuracy. 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.

Several observations can be made from Tables 7 and 8 regarding the performance of alternative data. First, Table 7 shows that, as one might expect for the non-linear algorithms there are noticeable differences between the predictive performance on training sets compared with corresponding the test sets suggesting they are overtraining more

than others. This is especially so for the neural networks and extreme Gradient Boosting where training AUC values reach 0.978 and 0.991, respectively. Second, extreme gradient boosting with oversampling or with principal components gives the highest predictive accuracy with AUC values of 0.625 and 0.624, respectively. Again comparisons with other papers are difficult because of the different variables in each author's models. However, given that our models have only alternative variables they perform well compared with those in other papers that include more conventional variables as well such as those by Agarwal et al. (2019) and Bojorkegren and Grissen (2018). In addition, Table 8 signals that email related alternative data are able to yield an even higher prediction performance on their own in some settings.

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 to 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.

6 Concluding remarks

Very few papers have used psychometric data to estimate credit scoring models, and it is rare for researchers to gain access to this type of data for individual borrowers. We know of no papers that have used email usage as a predictor of credit risk. One of the novelties of our work is that we have been able to gain data on email usage and psychometric characteristics of the same borrowers and to match these to the credit repayment performance of each borrower. The difficulty of gaining such data has inevitably constrained our sample sizes. Nevertheless, despite this, 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 email usage 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. Using these types of alternative data could help to reduce financial exclusion - the inability of individuals to gain credit because no risk score can be computed for them because using these variables, a score could be calculated. Our work suggests that since these types of variables increase predictive accuracy, it may be possible for lenders that have large amounts of conventional repayment data to have even more accurate models by using these variables than omitting them which is currently the case. More accurate PD models reduce bank risk and may enable more accurate risk-based pricing to be practised as well, although of course, the costs of gaining such data may be very high.

Appendix

A Complementary tables for descriptive analysis and literature review

Table A1: Descriptive statistics of the three datasets in Ensemble A.

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 (male=1)	0.499	1826
Income_cns_dol	987.25	1802
Psychometric		
Has accounts at other financial institutions	1.337	1751
Money now or in three months	1.5991	1826
Money now or in six months	1.6358	1826
Number of contacts	2.529	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
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	3.877	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	0.479	367
Log of number of emails received from uber.com	1.1709	442
Log of number of emails received from uber	1.1758	442

Authors	Alternative variables included	Sample	Results
Studies incorporating textual data			
Dorfleitener et al. (2016)	Aspects of textual description of purpose loan to be put to: spelling errors, length of text, types of keywords	Peer to peer loans	No predictive accuracy
Gao (2018)	Aspects of loan purpose descriptions: readability, tone, occurrence of deception cues	Prosper Peer to peer loans website	One sd reductiuon in readability, less positive tone, higher level of deception cues assocd with an increase in pd of up to 2.4%
Netzer et al. (2018)	Aspects of loan purpose and of credit applicant.	Prosper peer to peer webite	AUC increase of 2.64% when textaul characteristics added to financial and demographic characteristics. In crease greater for lower credit grades than for higgher grades.
Iyer et al. (2016)	Soft information e g whether borrower posts a picture, number of words used in listing	Prosper peer to peer website	Including soft information to financial variables and Experian credit score AUC increased from 0.710 to 0.714.
Berg et al. (2018)	"Digital footprint" variables: examples being whether used lower case when writing, email address errors operating system and device used	data from an e-commerce furniture company in Germany	Adding digital footprint variables to credit score increased AUC from 0.680 to 0.728for scorables and digital footprint varianles alone gave an AUC of 0.683 for unscorables.
Studies incorporating mobile phone data			
Bjorkegren & Grissen (2018)	Characteristics of mobile usage: measures of periodicity of usage, fraction of duration time spoken during a work day, variation in usage, autocorrlation between calls and SMS messages	Telecom loans in low-income per head country, data from EFL	Phone indicators gave higher AUC than credit bureau alone. When phone indictators added to bureau AUC increased 0.55 to 0.62.
Osbarsdottir et al. (2019)	Combined datasets base don call records, credit and debit details of customers to create social network information.		AUC increased 0.899 to 0.923 when added call records to tradional data
San Pedro et al. (2015)	Large number of characteristics of mobile phone use e.g. daily duraction of calls, daily SMS events	60k mobile phone customers in low income Latin American country.	AUC increases from 0.591 when credit buruea data alone is used to 0.675 when mobile data is added.
Tan et al. (2016)	Proportion of yearly visits to specific types of locations (from mobile phone locations),number of messages and calls to contacts, number of apps and sites visited per month.	Microloans fro a microfinance company located iin a lower middle income economy in South East Asia.	Adding location and network cohesion increased AUC from 0.736 to 0.763
Agarwal et al. (2019)	Number of apps, having specific installed apps such as dating, travel etc, Facebook status, LinkedIn status, large number of charateristics of calls such as duration of outgoing andnumber of missed	Loans up to around £2160 made by a large fintech in India	Adding apps and call characteristics increased AUC gained from credit bureau alone from 0.603 to 0.775 using random forest, but increased it from 0.483 to 0.492 using logistic regression.
Studies using Facebook data			
De Cnudde et al. (2019)	Characteristics of networks identified by content of Facebook messages that allow classification into individuals that are similar to each other, friends or friends that interact.	4512 users of Facebook and collected by Lenddo.	Inclusion of each type of network improves classification accuracy over socio demographics.
Lin et al. 2013	Levels of "friendship" depending on lending relationship.	205k listings on Prosper.com	Number of closer friends significantly affes=cts hazrad. No predictive accuracy statistics.
Social Network information			
Zhang et al.	Social network information: membership score, prestige, forum currency, contribution scores	20000 accounts for PPDai platform	When included with credit rating, loan information and otherse the accuracy was 81.2%. No indication of performance ontributed by network information.

Table A2: Summary of studies that use of alternative data to predict loan performance.

B Algorithms for Credit Scoring

B.1 Logistic regression

Logistic regression is one of the most popular methods used to analyse binary data (McCullagh and Nelder, 1989; James et al., 2013). Consider, for example, a sample of n cases and denote by y_i the indicator of default for case i. For each i, it is natural to assume that y_i follows a Bernoulli distribution with unknown parameter q_i, where q_i represents the probability of default for case i, $i = 1, ..., n$. These default probabilities can be estimated based on observable attributes. Typically, let us assume that m potential covariates have been observed. The dependence of q_i on these covariates is often expressed through the logit function as follows

$$\log\left(\frac{q_i}{1-q_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad i = 1, ..., n \tag{1}$$

In this expression, $\boldsymbol{x}_i$ is the known m-length vector of covariate values for account i, and $\boldsymbol{\beta} = [\beta_i, \cdots, \beta_m]^T$ is the vector of regression coefficients. The elements of $\boldsymbol{\beta}$ modulate the impact of the covariates on the default probabilities q_i. In practice, however, the true value of the vector of $\boldsymbol{\beta}$ is unknown. It is often estimated as the maximiser of the likelihood function L, given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} q_i^{y_i}(1-q_i)^{1-y_i} \tag{2}$$

B.2 Ridge regression

Ridge regression is a form of penalised regression. It allows one to prevent multicollinearity and to reduce model complexity using regularisation (Hastie et al., 2009; James et al., 2013). In ridge regression, the regression coefficients vector $\boldsymbol{\beta}$ is estimated as the maximiser of the penalised log-likelihood ℓ_p given by

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \lambda\, h(\boldsymbol{\beta}) \tag{3}$$

where $\ell(\boldsymbol{\beta})$ is the logarithm of the ordinarily likelihood function shown in (2), h is a L_2-norm regularisation function defined by

$$h(\boldsymbol{\beta}) = \sum_{r=1}^{m} \beta_r^2 \tag{4}$$

and λ is the regularisation parameter.

For a fixed value of the regularisation parameter λ, an estimate of $\hat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$ can be obtained by maximising the penalised likelihood (3). In general, λ controls the size of the

coefficients. For example, larger values of λ reduce the magnitude of resulting regression coefficients. The optimal value of the regularisation parameter can be selected via information criteria such as Ahaike Information Criteria (AIC) or Cross validation (CV); see Akaike (1974) or Craven and Wahba (1979). In our analysis, the optimal value of λ was selected via CV. For example, the curve of CV corresponding to the analysis of the alternative data from Ensemble B is shown on the left panel of Figure A1. The optimal value of λ is 0.134; the final regression parameters were estimated based on this value.

B.3 LASSO

LASSO (Least Absolute Shrinkage Selector Operator) is similar to ridge regression in the sense that complexity is simplified through regularisation (Hastie et al., 2009; James et al., 2013). With LASSO, the regression coefficients are estimated by maximising the penalised log-likelihood (3) but with a L_1-norm regularisation function instead, i.e.

$$h(\boldsymbol{\beta}) = \sum_{r=1}^{m} |\beta_r| \tag{5}$$

Unlike ridge regression, the LASSO regularisation function (5) shrinks the least important regression coefficients to zero. The larger the regulation parameter, the higher the number of coefficients shrunk to zero. The optimal value of the regularisation parameter can be chosen via Cross Validation. For example, in the analysis of the alternative data from Ensemble B in Section 4.1, with the optimal regulation parameter of 0.0026, only 11 variables (out of 237) were retained. The graph of the CV as a function of the regularisation parameter is shown on the right-hand side of Figure A1.

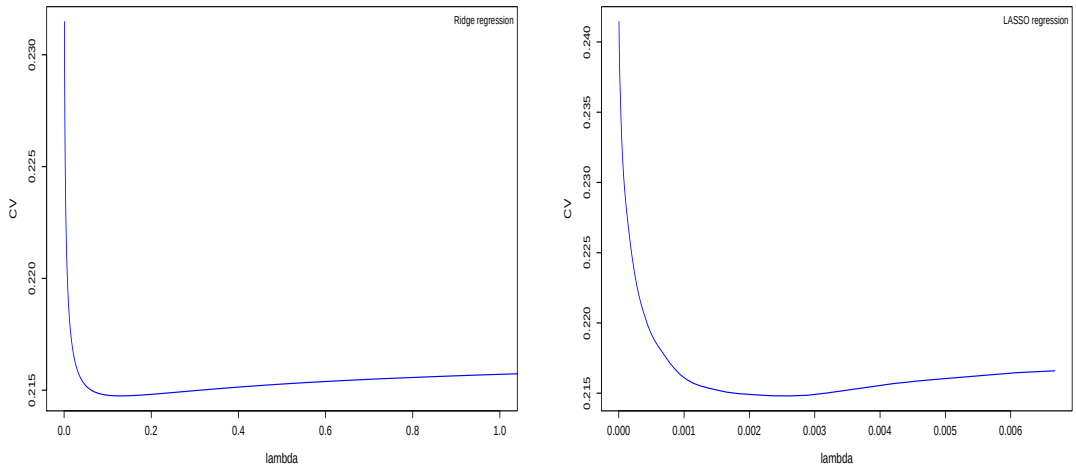
B.4 Gradient Boosting

Gradient boosting is a machine learning technique comprising an ensemble of learners built in a hierarchical fashion (Chen and Guestrin, 2016; Efron and Hastie, 2016). One of the most popular learners used in this context is regression trees. Thus, at each iteration of the hierarchy, a new tree is trained with respect to the error of the whole ensemble learnt so far, and then used to update the ensemble.

If $f(\boldsymbol{x}_i)$ is the prediction of y_i, let us denote by $\mathcal{D}(y_i, f(\boldsymbol{x}_i))$ the corresponding residual deviance, $i = 1, ..., n$. The generic gradient tree-boosting algorithm can be schematised as follows.

(i) Initialise the boosting model:

Figure A1: Optimisation of the regularisation parameter on the alternative dataset from Ensemble B.



- Set $f_0(\boldsymbol{x}_i) = \alpha$ where α is a real number and $i = 1, ..., n$.
- Estimate α as the minimiser of $\sum_{i=1}^{n} \mathcal{D}(y_i,\ f_0(\boldsymbol{x}_i))$.

(*ii*) Updates: for $k = 1, ..., K$, repeat

- Compute pseudo residuals: $r_i = \left[\dfrac{\partial \mathcal{D}(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{k-1}}, \quad i = 1, ..., n$
- Train a new regression tree $T(\boldsymbol{x})$ with respect to the pseudo residuals.
- Update the ensemble: $f_k(\boldsymbol{x}) = f_{k-1}(\boldsymbol{x}) + \delta\, T(\boldsymbol{x})$

(*iii*) Output the final boosting model: $\hat{f}(\boldsymbol{x}) = f_K(\boldsymbol{x})$

The performance of this algorithm is controlled by a number of parameters, including the depth K of the hierarchy, the complexity of the trees, and magnitude δ of the contribution of each tree. Selection of these parameters was carried out using a combination of grid search. For example, the performance of Gradient Boosting presented in Table 7 was achieved with $K = 2$ and $\delta = 0.41$.

B.5 Neural Network

Neural Network is a machine learning technique involving multiple *hidden* layers between the input data and the output (Goodfellow et al., 2016). It is typically represented by a network diagram as in Figure A2. The layers are made up of nodes, and that is where computation takes place. In general, a node combines inputs from the previous layer with a set of coefficients that either amplify or dampen the impact of the inputs.

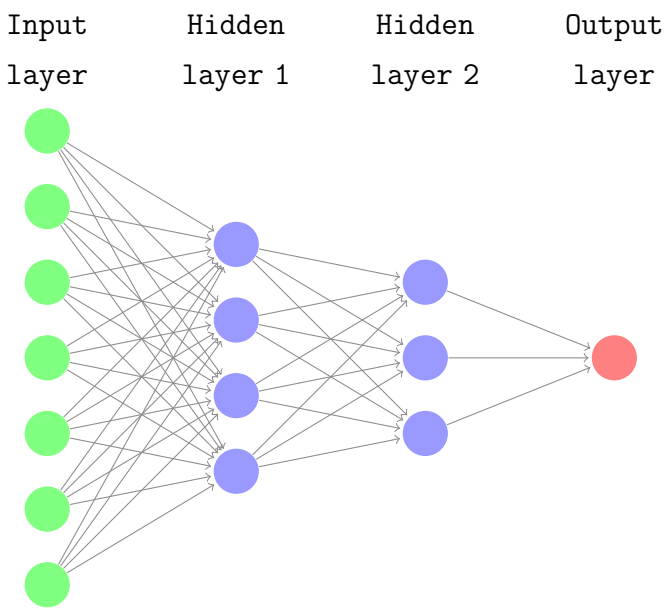


Figure A2: Graph of a fully connected neural network with an input layer (seven nodes), two hidden layers (four and three nodes) and one output layer (one node).

Let us consider a neural network with q layers. For a given account i, let us denote by $\boldsymbol{Z}_i^{\{l\}}$ the output vector from layer l, $l = 1, ..., q$. To train the network, these outputs are expressed recursively in terms of previous layers as follows

$$\boldsymbol{Z}_i^{\{l\}} = g\left(\boldsymbol{\Theta}^{\{l-1\}}\boldsymbol{Z}_i^{\{l-1\}} + \boldsymbol{b}^{\{l-1\}}\right), \quad \text{with } \boldsymbol{Z}_i^{\{1\}} = \boldsymbol{x}_i. \tag{6}$$

In this expression, $\boldsymbol{x}_i$ is the vector of covariate values associated with account i, $\boldsymbol{\Theta}^{\{l\}}$ is the matrix of weights associated with layer l, $\boldsymbol{b}^{\{l\}}$ is the vector of intercepts (often referred to as biases), and g is an activation function acting element-wise; that is: $g([a_1, ..., a_n]) = [g(a_1, ..., g(a_n)]$. The activation function can also be indexed by layers. Standard choices of activation functions include *sigmoid*, *arctan* and *radial basis* functions.

The matrices of weights $\boldsymbol{\Theta}^{\{l\}}$ and vectors of intercepts $\boldsymbol{b}^{\{l\}}$ shown in (6) are unknown. In practice, they are estimated iteratively and recursively by maximising an objective function through forward and backward propagations. For large networks, nevertheless, some regularisation is often imposed on the objective function, and this helps to improve the stability of the network. For binary response data, the objective function often used is similar to the logarithm of the likelihood (2). In particular, when modelling credit defaults via neural networks as in this section, the resulting outputs from the node in the final layer correspond to default probabilities.

The performance of a neural network depends on hyperparameters such as the number of layers, the number of nodes within layers, learning rate and activation functions. For

example, the prediction performance shown in 7 was obtained from a neural network with a sigmoid activation function and four hidden layers (40-29-20-12). This structure was obtained by optimisation of the objective function via a combination of grid search and random initialisations.

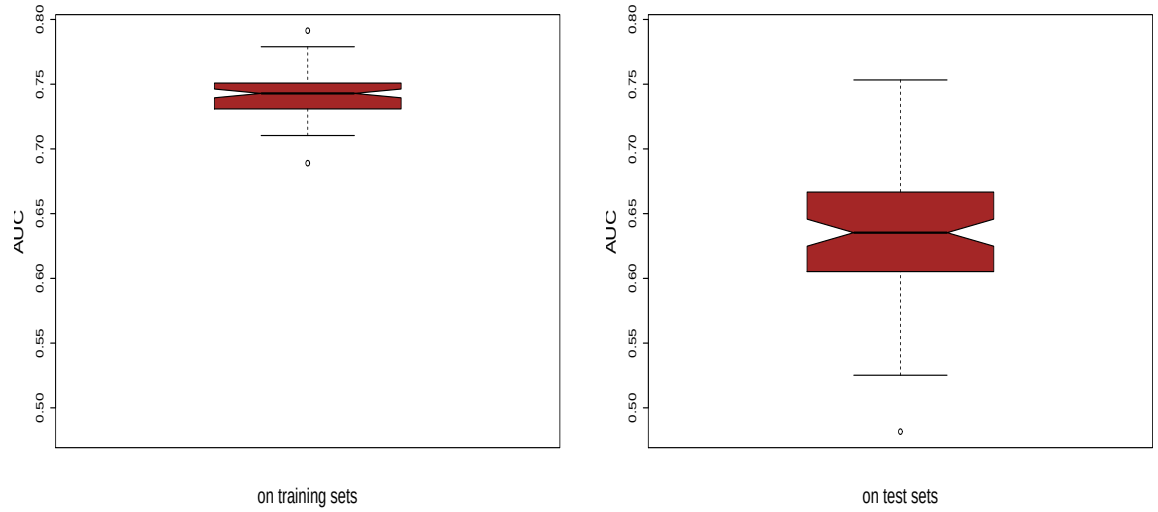
C Analysis of the datasets in Ensemble A using imputed values

The three datasets from Ensemble A introduced in Section 2 contain a substantial number of missing records. The second approach used to analyse these datasets in this paper was to impute missing values before estimating the scoring models. The starting point was to create a combined dataset by merging the three variable groups from Ensemble A using the id field (after removing rows with missing target value and filtering out low variance variables in each dataset, separately). Note that this combined dataset contains a substantial number of missing data for two main reasons. First, each contributing variable group has missing records, and second, the alternative variables were missing for a large proportion of cases in the combined dataset (indeed in the original alternative dataset, a valid target value was available on only 442 cases).

There are various methods in the literature to impute missing values, from simple mean/mode substitution through to more advanced imputation methods. The approach used in this analysis is the so-called multiple imputations 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 $(x_1, x_2, ..., x_p)$ and values are missing for some 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 observed values of this variable on observed and imputed values of all the other variables. Predict the missing values of x_2. Repeat this for all variables. Then repeat this 'cycle' a number of times (Royston and White, 2011).

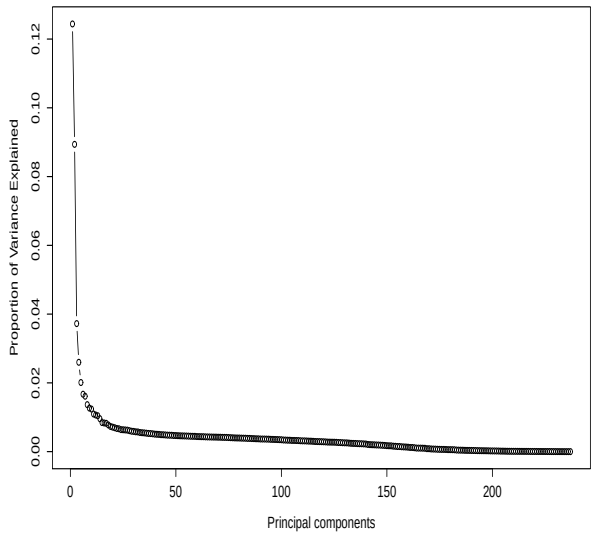
Using this imputation method, 20 completed datasets were generated based on the underlying patterns and uncertainty in the original data. On each of these datasets, a logistic regression model was estimated using the demographic variables alone. Afterwards, the 20 resulting models were averaged into one pooled demographic model following Little and Rubin (2002). Similarly, separate pooled psychometric and alternative models were

Figure A3: Prediction performance from the imputation based approach on the demographic characteristic, the psychometric variables and the alternative data.



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 3.

Figure A4: Relative importance of the principal components (Set B).



Bibliography

Agarwal S. and Alok S. and Ghosh P. and Gupta S. (2019) Financial inclusion and alternate credit scoring for the Millenials: role of big data and machine learning in Fintech. *Business School, National University of Singapore Working Paper*, SSRN 3507827.

Akaike H. (1974) A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Arraiz I. and Bruhn M. and Stucchi R. (2017) Psychometrics as a tool to improve credit information. *The World Bank Economic Review*, v30, Issue Supplemenet1, S67-S76.

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

Berg T. and Burg V. and 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.**.

Chen T. and Guestrin C. (2016) XGBoost: A Scalable Tree Boosting System. doi:10.1145/2939672.2939785

Carroll P. and Rehmani S. (2017) *Alternative Data and the Unbanked* Oliver Wyman Report.

Craven P. and Wahba G.(1979) Smoothing noisy data with spline functions. *Numerische Mathematik*, **19**, 377-403.

De Cnudde S. and Moeyersoms J. and Stankova M. and Tobback E. and Javaly V. and Martens S. (2019) What does your Facebook profile reveal about your credit-worthiness? Using alternative data for microfinance. *Journal of Operational Research Society*, **70 (3)**, 353-363.

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

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

Dlogosch T. J. and 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. and Priberny C. and Schuster S. and Stoiber J. and Weber M. and 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.

Efron B. and Hastie T. (2016) *Computer Age Statistical Inference*.Goa Q. and Lin M. and Sias R. (2018) Words matter: the role of texts in online credit markets. Cambridge University Press.*Available at SSRN 2446114*.

Friedman J. (2002) Stochastic gradient boosting. *Computational Statistics and Data Analysis*, **38**, 367-378.

Goodfellow I. and Bengio Y. and Courville A. (2016) *Deep Learning*. MIT Press.

Hand D. J. and Henley W. E. (1997) Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society A*, **160 (3)**, 523-541.

Hastie T. and Tibshirani R. and Friedman R. (2009) *The Elements of Statistical Learning*. Springer.

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

James G. and Witten D. and Hastie T. and Tibshirani R.(2013) *An Introduction to Statistical Learning*. Springer.

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

Klinger B. and 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. and Baesens B. and 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.*

Makela C. J. and Punjavat T. and Olson G. I. (1993) Consumers' credit cards and international students. *Journal of Consumer Studies and Home Economics*, **17**, 173-186.

McCullagh P. and Nelder J. A. (1989) *Generalised Linear Models*. Chapman & All/CRC.

Meier S. and Sprenger C. (2010) Present-Biased Preferences and Credit Card Borrowing. *American Economic Journal: Applied Economics*, **2(1)**, 193-210.

Netzer O. and 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.

Oskarsdottir M. and Bravo C. and Sarraute C. and Vanthienen J. and Baesens B. (2019) The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing Journal*, **74**, 26-39.

Raghunathan T. E. and Lepkowski J. M. and 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.

San Pedro J. and Prosperpio D. and Oliver N. (2015) Mobiscore: towads universal credit scoring from mobile phone data. In Ricci F. and Bontcheva K. and Coulan O. and Lawless S. (eds) *User Modelling, Adaptation and Personalisation*, 23rd International Conference, UMAP 2015, Dublin. Proceedings.

Schneider R. and Schutte A. (2007) The Predictive Value of Alternative Credit Scores. *Centre for Financial Services Innovation working paper.*

Tan T. and Bhattacharya P. and Phan T. (2016) Credit Worthiness prediction in micro-finance using mobile data: a sopatial-network approach. *Thirty Seventh International Conference on Information Systems*, Dublin.

Thomas L. C. (2000) A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16(2)**, 149-172.

Thomas L. C. and Crook, J. and Edelman, D. (2017) *Credit Scoring and Its Applications.* London: Siam.

Tong E. Mues C. and Thomas L. (2012) Mixture cure models in credit scoring: if and when borrowers default. *European Journal of Operational Research*, **218**, 132-139.