UNIVERSITY OF EDINBURGH
Business School

Credit scoring with alternative data

Jonathan Crook, Viani Djeundje, Raffaella Calabrese, Mona Hmid
Credit Research Centre, University of Edinburgh

Credit Scoring & Credit Control XVI Conference August 2019

UNIVERSITY OF EDINBURGH
Business School

Incidence of Unscorable People in US Federal Protection Bureau's CC Panel 2010

Brevoort, Grimm & Kambara (2016)

1-in-48 random sample from a NCRA
188.7m have credit records that can be scored by a commonly used credit scoring model.

Col	Population	No (m)	% of US Pop	Share with Performance	Delinquency Rate
1	Can be Scored	188.7	77.9	89.5	12.2
2	• Thick file	180.7		90.2	12.4
3	• Thin file	8.1	3.3	72.9	6.3
4	• Stale	9.6	4.0	12.3	26.0
5	• Insufficient	9.9	4.1	21.8	22.4
6=4+5	Cannot be scored	19.4	8.1		
7	**No credit records**	**26**	**10.7**		

UNIVERSITY OF EDINBURGH
Business School

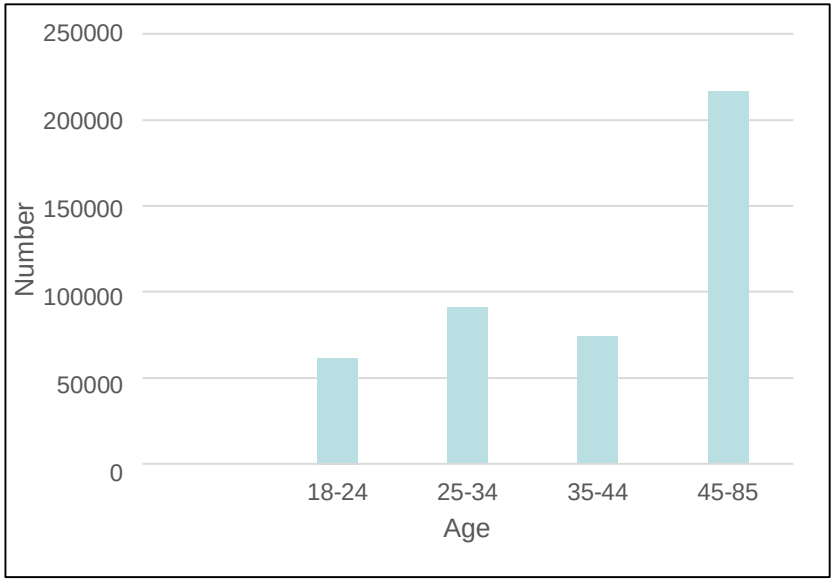
Incidence of Unscorable People in US: FICO Data 2015

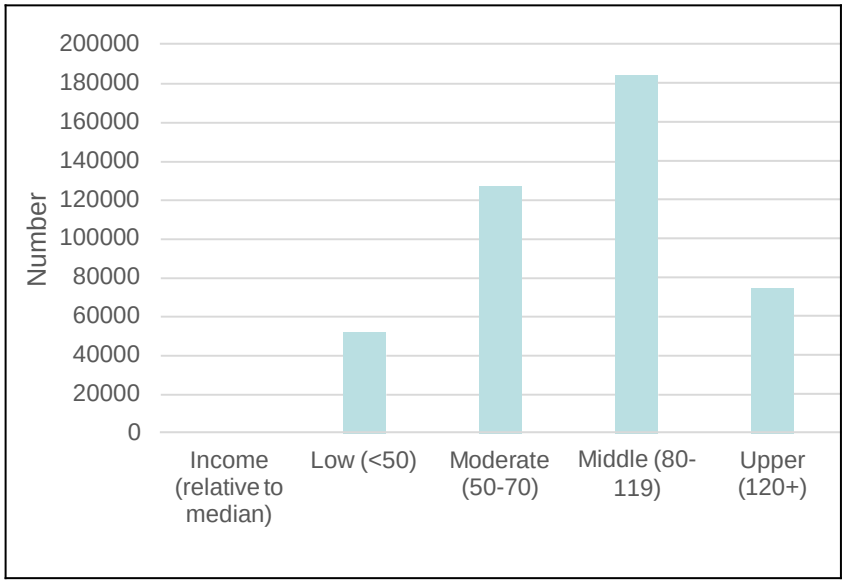
Segment		Millions
Have credit records but insufficient for scoring	No tradeline performance reported in last 6 months	3
	Files with delinquent tradelines, collections or adverse public records	18
	Stale files	7
No credit records		25
Total		**53**

Source: A. Jennings (2015) *Expanding the credit eligible population in the USA: A Case Study*. Credit Scoring and Credit Control XIV conference, Edinburgh.

Federal Reserve Bank of New York Consumer Credit Panel/Equifax

Sample: 15,509 individuals who had no credit score due to thin files or no credit file [lack of trade lines to construct score or no credit history] Q1 2007





Source: Smith, M. & Henderson, C. (2017) Beyond thin files. *Social Science Quarterly*, 1-19.

Worldwide

1.7billion estimated number of adults without an account at a financial institution nor facility through a mobile phone provider (31% of adult population) Demirguc-Kunt et al (2017)

Concentrated in lower income countries: China 204m, India 357m Indonesia 102m

A major difficulty is lack of past credit history.

Industry response: use alternative data.

Examples

- Rental data in UK (Experian)
- Data on utility poayments, evictions property values etc (FICO)
- Mobile phone data (CredoLab, Lenddo)
- Psychometrics (Lenddo,)

Contributions

Two contributions

1. To show that by using data on email usage and on psychometrics one can gain good separation between good and bad payers.

2. To show the relative contributions to predictive accuracy of these characteristics compared with demographic variables.

Literature

Textual predictors

Author	Predictors	Enhancement in predictive accuracy	Data
Dorfleitener et al (2016)	Spelling errors, length of text, occurrence of keywords in loan purpose statements		
Goa (2018)	Readability, tone, occurrence of deception cues in loan purpose statements		Prosper P2P
Netzer et al (2018)	Two word combinations, types of word groupings	Increase AUC by 2.6%	Prosper P2P
Iver et al (2016)	Soft information (whether posted a pic, number of words used in listing)	Increased AUC from 0.710 to 0.714	Prosper P2P
Berg et al (2018)	Digital foot print variables: use of lower case, email address errors , os of device etc	Increased AUC for scoreables: 0.68 to 0.73 Unscoreables gained 0.68.	E-commerce furniture seller

Mobile Phone predictors

Bjorkegren & Grissen (2018)	Periodicity of use, fraction of spoken time during the day, variation in usage, autocorrelation between calls and SMS messages	Phone indicators gave higher AUC than bureau alone.	EFL

Psychometric variables

Klinger et al (2013)	66 undefined variables	Used a lone: AUC 0.7	Micro & SMEs, Peru
Arraiz et al (2017)	undefined	Those accepted under traditional predictors and rejected on Psyc model had higher default rate than those accepted on a traditional model.	Banked entrepreneurs,
Dlogosch et al (2017)	Measures of conscientiousness, emotional stability, openness to experience and integrity.	N/a	Micro-entrepreneurs, Kenya
Liberati & Camillo (2018)	Six factors from a PCA applied to responses to Semiometrie	When added to models containing cash flow and solvency score, AUC increased 0.554 to 0.850.	

Data

Data set A

	Demographic	Psychometric	Email
# variables	12	350	53
# cases	1826	1826	33,091

Data set B

	Demographic	Psychometric	Email
# variables	n/a	n/a	237
# cases			16,358

Two stage estimation

- Logistic regression
- Stepwise using p-value or AIC
- One for each of:
  demographics, psychometrics, email

$\hat{y}_j$ $j=1, 2, 3$

Ensemble of separate models

$z = \sum_j b_j \hat{y}_j$

Logistic regression

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

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

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

UNIVERSITY OF EDINBURGH
Business School

Predictive Performance

Stage 1

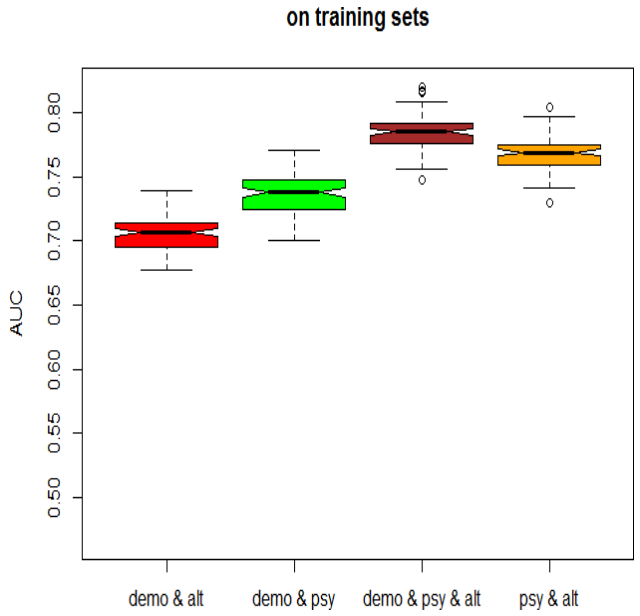
	Demographic model	Psychometric model	Email data model
AUC Training	0.63	0.67	0.67
AUC Test	0.62	0.60	0.58
# training cases	1370	1370	332
# parameters	11	23	13

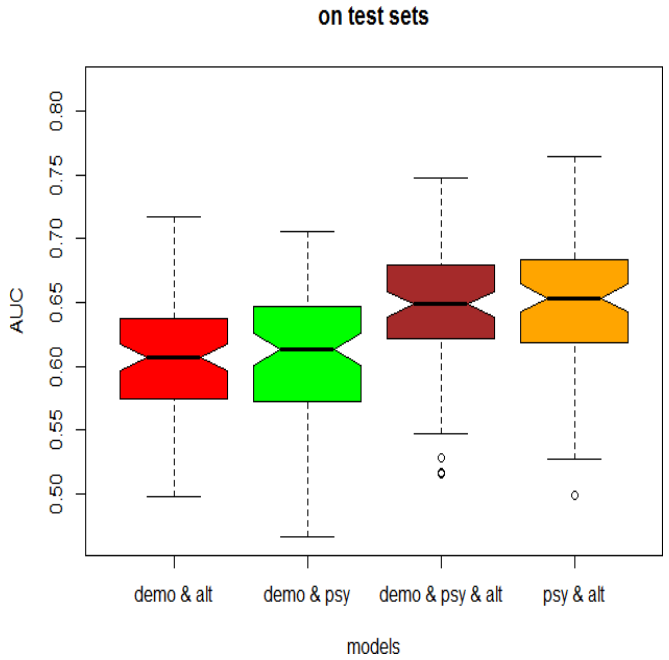
Stage 2

	Demographic + Psychometric	Demographic + Email	Psychometric + Email	Demographic + Psychometric + Email
AUC Training	0.66	0.71	0.71	0.73
AUC Test	0.75	0.69	0.67	0.72

Sensitivity with respect to train/test split

Results from 100 different random samples of 75% training 25% test.

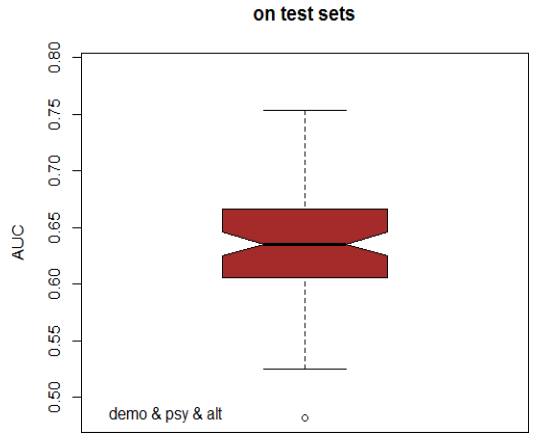




Imputation

To increase sample size we pooled all 3 datasets and imputed missing values.

- Used MICE (multiple imputations by chained equations) Royston & White 2002. 20 implicates.
- Estimate for each implicate and average using formula from Little & Rubin (2002).
- Estimate separate models for each of demographic variables, psychometric and email datasets.
- Ensemble.



Results similar to the previous analysis.

Analysis of Data set B

	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%

UNIVERSITY OF EDINBURGH
Business School

Discussion

- Predictive accuracy from ensemble model greater than that from individual models.

- AUC gained from our models compares well to others in the literature that use
- alternative data:

Our models (demographic, psychometric, email)	0.75
Berg (digital footprint)	0.73
Dlogisch (psychometric)	0.67
Liberati & Camillo (psychometric)	0.85

Implications

- Use of alternative data is increasing in lower income countries
- Can gain commercially useful predictive accuracy using psychometric and/or email variables
- Can be useful substitute for lack of credit performance data
- But applicants must be willing to complete a psychometric test before applying and possible gaming