Levels of Over-Indebtedness in the UK

Technical Report
2018 Over-Indebtedness Model

Money & Pensions Service

CACI

Table of Contents

1. Foreword **3**

2. Background **4**

3. Data Sources **5**

3.1. Research Data 5

3.2. CACI Ocean Data 6

4. Defining "Over-Indebtedness" **8**

5. Cleaning/De-duping Respondents and Tagging with Ocean Data **9**

6. Modelling Process **11**

6.1. 2015 Over-Indebtedness Model 11

6.2. Model Update Approach 11

6.2.1. Pass/ fail criteria: Existing Model Validation 12

6.2.2. Pass/ fail criteria: Model Refresh 12

6.2.3. Pass/ fail criteria: Model Rebuild 12

7. Model Parameters **14**

7.1. Previous models, 2015-2017 14

7.2. The 2018 model 14

7.2.1. Existing Model Validation 14

7.2.2. Model Refresh 16

7.2.3. Model Rebuild 17

7.3. Variable Definitions 19

8. Evaluation of Model **21**

8.1. Statistical Significance of Parameters 21

8.2. Hosmer-Lemeshow Test 21

8.3. *C*-Statistic 22

8.4. Multicollinearity 22

8.5. Model Validation 23

8.6. Stability of Model and Prediction 24

9. Contacts **24**

1. Foreword

The purpose of the Money and Pensions Service is clear: to help create 'a society where everyone makes the most of their money and pensions'.

One of the biggest problems we face in pursuing this goal is the large proportion of people in the UK who continue to struggle with debt. Around nine million are now over-indebted and the supply of debt advice services is still insufficient to meet demand.

This challenge means it is critically important to make effective decisions, that make the best use of all available resources. To do this, we need intelligence. We need to understand:

- **How many** people are over-indebted?
- **Where** are levels of over-indebtedness highest?
- **Which groups** of people are most likely to be over-indebted?
- How is over-indebtedness **changing over time?**

Our over-indebtedness model, produced and updated by CACI using consistent trend data since 2015, helps us to answer these questions by estimating the level of over-indebtedness in every country, region, local authority and constituency, and for specific demographic groups.

The 2018 results can be found at: https://www.moneyadviceservice.org.uk/en/corporate/a-picture-of-over-indebtedness-in-the-uk.

It is critical that decisions are based on robust data. This technical report explains the full methodology for updating the model to produce our 2018 estimates, including the results of statistical acceptance tests. If you have any other questions on the model, you can get in touch with any of the contacts listed at the end of this report.

Sarah Little
Insight Manager
Money and Pensions Service
Email: sarah.little@maps.org.uk

2. Background

The Money Advice Service, one of the legacy organisations that formed the Money and Pensions Service, has been measuring individuals' levels of over-indebtedness since 2012.

CACI first worked with The Money Advice Service in 2015 to produce a nationwide model, which combined large numbers of survey respondents with CACI's rich consumer data and resulted in over-indebtedness estimates for a range of geographies. The approach was a "bottom-up" methodology, meaning individuals were modelled separately and then aggregated into geographies based on their residential postcode.

In subsequent years, The Money Advice Service collected additional and updated research data, which CACI used to test, validate, and where necessary update the over-indebtedness model. This ensures a current and robust view of existing levels of over-indebtedness, as well as the predicting factors and the characteristics of the over-indebted population.

The 2016 and 2017 models were a "refresh" of the original 2015 model – meaning that only minor changes to the variables and parameters were required to produce over-indebtedness models that performed well. However, in 2018 a more comprehensive model rebuild was required to reach a model that satisfied the accuracy and robustness criteria. This involved investigating new variables for inclusion and resulted in a model that not only performs better than the previous year's model, but reflects the up-to-date social environment. Work patterns and economic climate are dynamic factors in the context of over-indebtedness, and it is important to re-assess these and ensure that a model represent the real-world situations individuals currently face.

It is important not only for the solution, but also the annual update process, to be clear and understandable, taking a transparent approach to the way over-indebtedness is calculated on an annual basis. This report summarises the original approach and details this year's changes to the Over-Indebtedness Model. The results are published on the Money and Pensions Service website[1].

[1] moneyandpensionsservice.org.uk/research/ or www.moneyadviceservice.org.uk/en/corporate/a-picture-of-over-indebtedness-in-the-uk

3. Data Sources

3.1. Research Data

The Money Advice Service provided CACI with research survey data for analysis. The total sample size was approximately 20,000 individual respondents.

This data was obtained from a survey undertaken with UK adults aged 18+ in June – August 2018.

Consistent with the data used in 2016 and 2017, and the majority of 2015 data used in the seminal model, the survey was undertaken via online panel.

The interviews were conducted online by Alligator (BDRC), using panels Research Now/ SSI and Panelbase. Respondents were de-duplicated between panels.

Blending several online panels is considered methodologically valuable as it limits the impact of any bias that may exist in an individual online panel, and within tracking studies can mitigate the impact of changes that panel providers may make to their panel year-on-year. The panels chosen to blend were selected for their size and reputation (with Research Now/ SSI being one of the largest and most respected panel providers in the UK) as well as the diversity of their recruitment methods. Large size and varied recruitment both reduce the risk of panel bias.

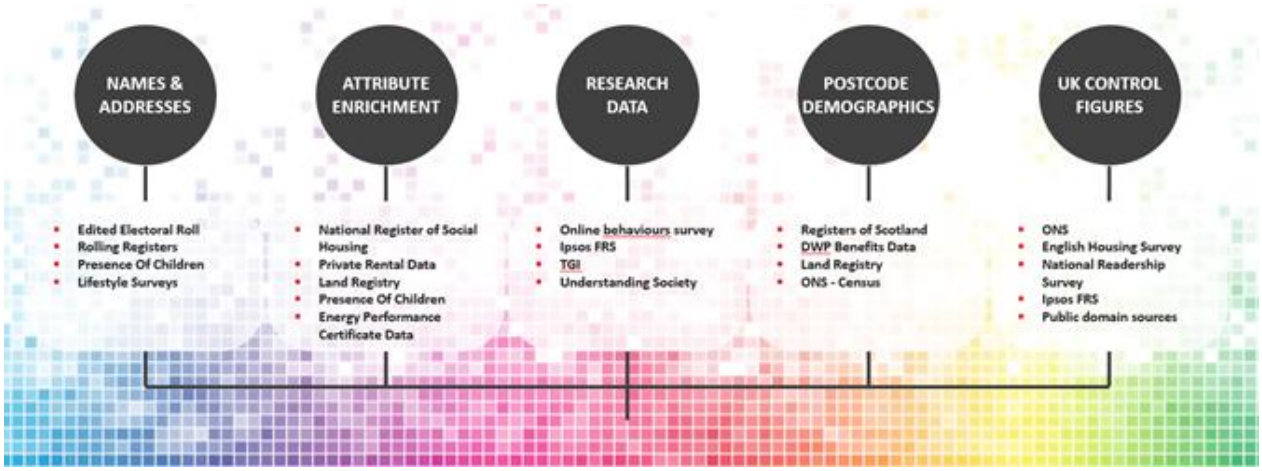
Quotas were set on region, age, gender and social grade to ensure that the input sample was sufficiently regionally and demographically representative to use within the modelling process (see "6.2 Model update approach" later in this document).

The sample size from each of the three panels was as follows:

Conducted by	Sample Size
Research Now	14,810
SSI	2,934
Panelbase	2,849

3.2. CACI Ocean Data

Ocean was used to build the original over-indebtedness model in 2015, and then used each year to validate and refine the model. It is an attribute-rich consumer database for the UK, maintained by CACI and updated quarterly. Hundreds of millions of records from research surveys, open data, government data and many other sources are collated together to create the universe.



Ocean includes:

- 48.9 million adults allocated to addresses, of which names are available for three-quarters of individuals. In order to ensure GDPR-compliancy, some records of individuals have been removed across the last few years, resulting in fewer names being available for individual-level matching. However, household-level matching remains unaffected.
- The name and address base forms the 'spine' of the Ocean database. It is built by merging and de-duplicating names and addresses from multiple high-volume sources, and selecting the most up to date information.
- A wide range of variables for each individual. Values are inferred from modelling based on other known characteristics taken from multiple sources. These sources include Land Registry information, Target Group Index (Kantar) and – particularly for those variables featuring in the over-indebtedness model – the widely-respected Financial Research Survey (Ipsos).
- Modelled estimates can be provided as categorical assignments for appropriate variables such as tenure, as inferred Yes/No flags, or as probability estimates that a person has an attribute – it is these latter propensity scores that are used in over-indebtedness modelling.

The real and modelled variables on Ocean cover a wide range of attributes, attitudes and behaviour. They include:

ATTRIBUTES

Age and gender
Number and age of children
Household Income
Household size and composition
Housing: type, tenure, size, value
Occupation
Social Grade
Number, age and type of cars

ATTITUDES

Adoption and engagement with technology
Attitudes to financial products and channels
Intention to switch financial products
Attitudes to online privacy and safety
Lifestyle attitudes
Shopping attitudes
Attitudes to the environment

FINANCIAL BEHAVIOUR

Financial products owned
Savings and Investments value
Credit card patterns of use
Loans and debt
Channel preference
Medical insurance

LIFESTYLE

Technology ownership and use
Holidays: destination, type, spend and booking method
News and Magazine readership
Interests and hobbies
Internet usage: frequency, location and technology
Types of goods and services purchased online
Online activities: gambling, dating, gaming etc.
Social networking: which networks and types of activity
Mobile phone: type of phone and how used
Shopping: types of stores visited (premium, mass, value)

4. Defining "Over-Indebtedness"

OVER-INDEBTED
Finds meeting monthly commitments a heavy burden and/or regularly in arrears with bills

The Money Advice Service first investigated the characteristics of over-indebtedness in the 2013 research, "Indebted Lives: the complexities of life in debt"[2]. The definition of over-indebtedness has remained consistent since then, and the component questions have been asked on research surveys each year. Over-indebted individuals are those that answer either:

i. I find keeping up with bills and credit commitments a **heavy burden**
ii. I have fallen behind on, or missed payments in **three or more months** out of the last six months

Note that the three months in (ii) do not need to be consecutive. Individuals may respond positively to one or both of the above questions to be identified as over-indebted.[3] Those that do not respond positively to either question are defined as "not over-indebted".

These questions feed into a single "Yes/No" binary variable that is modelled to predict over-indebtedness at an individual level.

Within the **raw unweighted** data supplied, the average proportion of respondents finding bills a heavy burden was 11.3%, while 12.2% of respondents had been in arrears in three of the last six months.

The table below shows how these figures vary across the three different survey sources.

Survey Source	Respondents	Keeping up is a heavy burden (unweighted %)	Arrears in 3 of last 6m (unweighted %)	Over-Indebted (unweighted %)
Research Now	14,810	10.73%	12.08%	18.09%
SSI	2,934	12.78%	12.47%	20.07%
Panelbase	2,849	12.81%	12.29%	19.66%
Total	**20,593**	**11.31%**	**12.16%**	**18.59%**

The figures in this table are unweighted calculations, made before the data was cleansed and matched to CACI records. The over-indebtedness figures above are therefore not representative of the UK population. They are however included here to show reported levels within the input survey data, and to demonstrate how these levels were relatively consistent across the three sources.

[2] https://www.moneyadviceservice.org.uk/en/corporate/indebted-lives-the-complexities-of-life-in-debt

[3] "To what extent do you feel that keeping up with your bills and credit commitments is a burden?" [A heavy burden; Somewhat of a burden; Not a burden at all; Don't know].

"In the last 6 months, have you fallen behind on, or missed, any payments for credit commitments or domestic bills for any 3 or more months? These 3 months don't necessarily have to be consecutive." [Yes; No ; Don't know]

5. Cleaning/De-duping Respondents and Tagging with Ocean Data

The first stage of the model update was to match the survey respondents to CACI's database of individuals. This appended the Ocean attributes and characteristics to each respondent, so that the model variables could be retested and validated against the dependent over-indebtedness variable derived from the research.

This stage also matches respondents to an individual within the UK consumer database, which allows for accurate de-duping amongst the surveys and waves. 79 records were removed as they appeared more than once across the research panels, and a further 464 records did not match to CACI Ocean universe, and therefore could not be used in the modelling process.

Additionally, the decision was made to remove a further 1,181 respondents, which came from a specific Research Now sub-panel. During the initial model validation within phase one, it was evident that these individuals demonstrated much higher levels of over-indebtedness than other respondents, even when demographics and regionality were taken into account. So as to avoid introducing bias into the modelling procedure these respondents were discarded as outliers before moving onto the model refresh stage in phase two.

A comment should be made on the proportion of postcode matches, which is higher than in previous years. This is because of changes brought about by GDPR, which meant that at the time of collection, panels were less likely to hold *detailed* up-to-date personal data or obtain consent to share it externally. As a result, there were fewer individual-/household-level matches to the Ocean database, and for those records we needed to use postcode-level data.

As part of GDPR preparations, tests were run on a sample of respondents with full personal data to gauge the impact of lower levels of matching. Reassuringly there were no significant differences in the attribution of variables and the resultant models between individual/ household and postcode-level matching. Additionally, within the subsequent modelling process, no differences in accuracy of prediction were seen between Panelbase (where all the sample was postcode only) and Research Now/ SSI respondents (where 71% of the sample matched at individual or household level).

Survey Source	*Total Records*	Individual or Household Match	Postcode Match	Duplicate Respondents	Unmatched to Ocean	Outlier Respondents *
Research Now	*14,810*	11,510	1,666	72	381	1,181
SSI	*2,934*	1,065	1,833	7	29	0
Panelbase	*2,849*	0	2,795	0	54	0
Total Proportion		**61%**	**31%**	**<1%**	**2%**	**6%**

*Outlier respondents were removed between phases one and two – see 7.2.1 for further detail

After matching to the CACI data universe, de-duping and removing outliers, the total sample size of usable records used to create the 2018 model was **18,869**.

This was split into a training sample (n=15,095) and a 20% validation sample (n=3,774) – the latter to independently verify a revised model using respondents that haven't contributed to the recalibration of

its parameters. This type of validation sampling – often referred to as "out-of-sample" validation – is considered statistical best practice in predictive modelling.

Care was taken to ensure the training and validation samples are representative of each other. Sampling was conducted using a "1 in 5" method on respondents stratified by research panel, demographic segment, region, and match level.

The sample sizes for each of the regions of the UK were as follows:

Region	Sample Size (Training Sample)	Sample Size (Validation Sample)
North East	608	156
North West	1,653	410
Yorkshire and The Humber	1,249	311
East Midlands	1,149	288
West Midlands	1,250	306
East of England	1,517	386
London	1,786	446
South East	1,925	481
South West	1,355	341
Wales	782	191
Scotland	1,359	340
Northern Ireland	462	118
UK Total	**15,095**	**3,774**

6. Modelling Process

6.1. 2015 Over-Indebtedness Model

In 2015 CACI worked with the Money Advice Service to produce estimates of over-indebtedness for the UK, and for each local authority. This was based on a logistic regression analysis of 11,279 survey respondents, which modelled each individual's likelihood of being over-indebted. The resulting model consisted of sixteen variables, and the report on this work is still available on the Money Advice Service website[4].

In subsequent years, this model has been re-tested and validated against renewed research data.

6.2. Model Update Approach

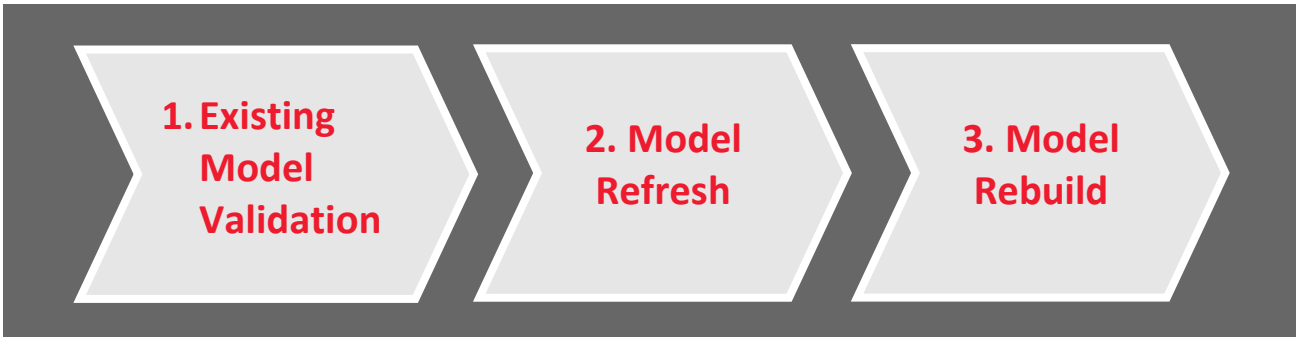
New research is commissioned each year to collect up-to-date data on levels of over-indebtedness within the United Kingdom. In 2016, 2017 and 2018 the research came from online panels, and the sample sizes were in the region of 20,000.

New survey data should be demographically and regionally representative of the UK population, and quotas are set to ensure this (see "3.1 Research Data" above). Some minor variations can be permitted as the model is built at individual level (and then aggregated to small areas), and CACI population data is able to take this into account. The model takes each respondent's region and demographic into account when calculating their own individual-level likelihood of being over-indebted. This model – including demographic and regional factors – is then applied to all adults in the UK, which accounts for any differences between the make-up of the survey respondents and the UK population.

Each adult is represented within Ocean, and therefore weighting is not required. However large variations in the sample may indicate bias, and sufficient volumes of each region characteristic are required to ensure statistical significance.

This data is run through a three-step validation process which tests the suitability of the current over-indebtedness model:

1. Existing Model Validation
2. Model "Refresh"
3. Model "Rebuild"



[4] www.moneyadviceservice.org.uk/en/corporate/a-picture-of-over-indebtedness-in-the-uk

Pass/ fail criteria are set up for each step, and a model is accepted if all of the criteria within a step are passed. If a step fails, the next step is undertaken. The pass/ fail criteria for the three phases are detailed below.

In both 2016 and 2017, the existing model validation failed and a model refresh (step two) was required. This year, in 2018, a model refresh failed to reach an acceptable model and a model rebuild (step three) was required.

6.2.1. Pass/ fail criteria: Existing Model Validation

The existing over-indebtedness model – i.e. its variables, parameters and intercepts – is applied to the latest survey data. The accuracy of the model is acceptable when the following criteria are met:

- When the existing model is applied to the sample, the predicted level of over-indebtedness is sufficiently close to the observed level amongst the respondents. An error of one standard deviation is permitted (which in practice means an acceptable error of around +/- one percentage point).
- The concordance statistic (*c*, a measure of individual-level accuracy) is sufficiently close to that of previous models. Given the 2016 and 2017 model refreshes returned 69% and 71%, *c* should be greater than or equal to 70%. However *c*>65% is permitted if the model performs well elsewhere.
- The fitted model passes Hosmer-Lemeshow's "Goodness-of-Fit" test. As in previous years this is assessed by analysing ten deciles and showing there is no evidence to support a lack of fit, with *p* – the probability of finding such lack of fit by chance – greater than or equal to 0.25.

6.2.2. Pass/ fail criteria: Model Refresh

If one or more of the above conditions are not met, then minor adjustments are made to the over-indebtedness model to produce improvements in accuracy or robustness. This is achieved by applying a logistic regression to the new research data, forcing the same variables into the model to elicit new parameters and intercept. Additionally, variables may be removed if statistically insignificant, and variables from previous years' models may be tested in the model for improvements. Note that *new* variables are not tested until the third step, the model rebuild. A model is accepted at this stage if the following criteria are met:

- The concordance statistic (*c*, a measure of individual-level accuracy) is sufficiently close to that of previous models. Given the 2016 and 2017 model refreshes returned 69% and 71%, *c* should be greater than or equal to 70%. However *c*>65% is permitted if the model performs well elsewhere.
- The fitted model passes Hosmer-Lemeshow's "Goodness-of-Fit" test. As in previous years this is assessed by analysing ten deciles and showing there is no evidence to support a lack of fit, with *p* – the probability of finding such lack of fit by chance – greater than or equal to 0.25.
- All variables retained in the model, as well as the intercept, are statistically significant at a level greater than 95% confidence.
- All variables retained in the model act in the same direction as previous years (i.e. positive and negative predictors remain positive and negative).
- A 20% validation sample (obtained by a stratified 1-in-5 sampling process) produces a modelled over-indebtedness within 10% of the reported level. Additionally the *c*-statistic is at least 65% within this sample.

6.2.3. Pass/ fail criteria: Model Rebuild

If one or more of the above conditions are not met, then full statistical analysis is run on the new research data in order to build a new logistic regression model. In this case, additional variables are sought, and new interactions tested.

A model is accepted at this stage if the following criteria are met:

- The concordance statistic (*c*, a measure of individual-level accuracy) is sufficiently close to that of previous models. Given the 2016 and 2017 model refreshes returned 69% and 71%, *c* should be greater than or equal to 70%. However *c*>65% is permitted if the model performs well elsewhere.
- The fitted model passes Hosmer-Lemeshow's "Goodness-of-Fit" test. As in previous years this is assessed by analysing ten deciles and showing there is no evidence to support a lack of fit, with *p* – the probability of finding such lack of fit by chance – greater than or equal to 0.25.
- All variables retained in the model, as well as the intercept, are statistically significant at a level greater than 95% confidence.
- All variables in the model pass the "common sense" test and appear a reasonable predictor of over-indebtedness. Where possible contradictions with previous years' models should be avoided.
- A 20% validation sample (obtained by a stratified 1-in-5 sampling process) produces a modelled over-indebtedness within 10% of the reported level. Additionally the *c*-statistic is at least 65% within this sample.

It should be noted that this stage was reached for the first time in 2018. Despite this, within the final model there are many overlaps with previous years in terms of model variables, and both accuracy and robustness are comparable to that of the original model in 2015.

7. Model Parameters

7.1. Previous models, 2015-2017

The 2015 over-indebtedness model consisted of sixteen variables, some of which were Ocean variables and some of which were combinations and interactions of variables. Twelve were positive factors (suggesting an increased likelihood of over-indebtedness), and four were negative factors (suggesting a decreased likelihood of over-indebtedness).

The model update in 2016 tested these sixteen variables against new research data. The 2015 model did not satisfy the statistical criteria required – five of the variables were found to be no longer statistically significant and were therefore removed from the model. The remaining eleven parameters (and the intercept) were adjusted accordingly.

There were few changes required to the model between 2016 and 2017. Aside from parameter and intercept adjustments, only the Scotland indicator variable was removed from the model.

The table below summarises these changes:

Parameter	2015 Parameter Co-efficients	2016 Parameter Co-efficients	2017 Parameter Co-efficients
Intercept	*-1.925*	*-2.205*	*-1.850*
Has Loan for Consolidation	4.584	5.808	7.449
Private Renting	0.315	0.365	0.446
Social Renting	0.431	0.255	0.410
Has 3+ Children	1.050	1.159	1.065
Single Parent	0.209	-	-
Social Grade D or E	1.067	1.770	0.873
Northern Ireland	0.527	-	-
Value of Home <£100k, South East	0.831	-	-
Value of Home <£100k, London	4.464	-	-
Unemployed, Wales & West Midlands	1.952	-	-
Household Income <£10k, Household Size 3+	1.159	1.578	1.552
Own Home Outright, Wales	0.670	0.369	0.532
Has Savings £10k+	-2.127	-1.933	-2.362
Aged 65-74	-0.919	-0.809	-0.916
Aged 75+	-1.211	-1.012	-1.071
Scotland	-0.259	-0.210	-

All coefficients are statistically significant at a 95% confidence level.

7.2. The 2018 model

7.2.1. Existing Model Validation

In 2017, the over-indebtedness model was built and calibrated to a UK figure of 15.8%, which was the reported and unweighted level within the training sample derived from the 2017 research.

In contrast, the 2018 survey data taken into phase 1 (after matching and de-duplicating) presented a reported and unweighted level of over-indebtedness of 18.1%, a figure 2.3 percentage points higher than the previous year.

When the 2017 over-indebtedness model was applied to this data, the modelled level of over-indebtedness was 16.7%. Despite the concordance statistic (a measure of individual-level predictability) being 72.1%, the overall forecast produced by the model was not at the required level. The over-indebtedness forecast was 1.4 percentage points lower than that reported by the respondents and outside of the acceptable levels of error, specifically one standard deviation.

	2017 Model applied to 2017 Respondents	2017 Model applied to 2018 Respondents
Modelled Over-indebtedness:	15.8%	16.7%
Observed Over-indebtedness:	15.8%	18.1%
Modelled Error (%)	n/a	-7.9%
Modelled Error (pp)	**n/a**	**-1.4pp**
Concordance Statistic (c)	**70.9%**	**72.1%**

Therefore, the decision was made to reject the 2017 model, and move onto the second phase of recalibrating the model with adjusted parameters.

Within the investigative analysis carried out, it was apparent that there was a group of respondents that consistently reported higher levels of over-indebtedness – even when contributing factors such as broad demographic segment and region were taken into account. These 1,181 respondents all came from a single sub-panel, and the decision was taken that these were anomalous records that should be removed from further analysis.

This reduced the number of records to 18,869, and the level of reported over-indebtedness from 18.1% to 17.0%. Details of the subsequent training and validation samples taken forward into phase 2 are shown in the table below.

Sample taken into phase 2	*Total Records*	Over-Indebtedness (%)
Training Sample (80%) -used to build/calibrate model	*15,095*	17.1%
Validation Sample (20%) -used to independently verify model	*3,774*	16.8%
Total	***18,869***	**17.0%**

7.2.2. Model Refresh

Running logistic regression with the ten variables derived in 2017 on the 2018 research data (a 20% sample was reserved for validation testing, see section 8.5) suggested new parameters and intercept.

Parameter	Model Parameters	Pr > Chi Sq
Intercept	*-1.495*	*<.0001*
Has Loan for Consolidation	2.990	**0.0594**
Private Renting	1.389	<.0001
Social Renting	1.099	<.0001
Has 3+ Children	1.224	<.0001
Social Grade D or E	**-0.755**	0.0068
Household Income <£10k, Household Size 3+	2.839	<.0001
Own Home Outright, Wales	-0.023	**0.9380**
Has Savings £10k+	-3.145	<.0001
Aged 65-74	-1.008	<.0001
Aged 75+	-1.137	<.0001

Two variables became statistically insignificant: the likelihood of having a loan for consolidation purposes, and the likelihood of owning your home outright (applied to residents of Wales only).

The parameter on the latter decreased close to zero, and in fact changed direction compared to 2017. The implication is that, with the large confidence band required around this parameter, we cannot even be sure (within 95% certainty) whether this variable has a positive or negative effect on over-indebtedness. This variable, "Own Home Outright, Wales" therefore could not be retained within a 2018 model.

The parameter on "has loan for consolidation" was only marginally insignificant at the 95% level (p=0.0594), and could therefore be retained and observed within alternative model adjustments.

It should be noted that "Social Grade D or E" remained statistically significant, but its direction changed from having a positive relationship with over-indebtedness to having a negative relationship. This suggests that those in lower social grades (unskilled manual, casual workers, unemployed etc) are less likely to be over-indebted. This contradicts previous years' models, and is likely the result of the model making small adjustments alongside a correlated variable or set of variables.

This second phase, model refresh, also allowed for additional categorical variables to be tested in the model where one or more variables already feature – namely region and age band. When this took place, two additional regional variables (London and Scotland) and one additional age variable (25-34) entered the model, and the parameter estimates are shown in the table below.

Both "London" and "Aged 25-34" acted as positive factors and "Scotland" as a negative factor of over-indebtedness. This tallies with characteristics seen in previous years, as well as corresponding with current beliefs on debt. It should be noted that "Wales" did *not* enter the model as an additional regional factor, and therefore the model did not require a replacement or regional proxy for the "Own Home Outright, Wales" variable removed.

All variables were now statistically significant at the 95% confidence limit. Although "Social Grade D or E" was significant (albeit marginally at 95% level, p=0.0441), this variable also retained its change in direction from 2017. This fails one of the model acceptance criteria and so this variable was also removed from the model.

Parameter	Adjusted Model Parameters	Pr > Chi Sq
Intercept	*-1.656*	*<.0001*
Has Loan for Consolidation	-	
Private Renting	1.168	<.0001
Social Renting	0.886	<.0001
Has 3+ Children	1.081	0.0001
Social Grade D or E	-	
Household Income <£10k, Household Size 3+	1.621	0.0212
London	0.189	0.0046
Aged 25-34	0.210	0.0003
Own Home Outright, Wales	-	
Has Savings £10k+	-2.454	<.0001
Aged 65-74	-1.184	<.0001
Aged 75+	-1.477	<.0001
Scotland	-0.248	0.0064

"Has Loan for Consolidation", "Social Grade D or E" and "Own Home Outright, Wales" have been removed, and additional age and region variables tested.

In terms of model accuracy, the concordance score was good (73.4%), but the model failed the Hosmer-Lemeshow goodness-of-fit test with a p-value of just 0.030 (the acceptance criteria demands p>0.25).

Therefore the decision was made to reject the "refreshed" 2017 model, and move onto the third and final step of rebuilding the model with new and revised variables for 2018.

7.2.3. Model Rebuild

Since adjusting the parameters of the 2017 over-indebtedness model did not produce a satisfactory model, the update process continued to a full rebuild. This allowed for all possible variables and interactions to be investigated, which resulted in new factors entering the model. However, where possible a new model and its variables should be similar to that of previous years. This maintains a level of consistency and helps prevent large step-changes in year-on-year predictions. A development of the 2017 model was preferred to a new model altogether.

Additionally, and more importantly, the new 2018 model needed to pass the accuracy and robustness criteria set out for other model update phases, and indeed the strict acceptance criteria first established for the original model in 2015.

An alternative model was established, which contained ten variables. Five of these variables exist in the 2017 model and the remainder can be thought of as close substitutes in most cases. Full variable statistics for this model are given in the table below. The sign of the parameter coefficients indicates whether the variable has a positive or negative effect on over indebtedness. To understand how much each variable affects over-indebtedness estimates, we need to look at the marginal probabilities. Presented in the last column of the following table, the average marginal probability describes how the likelihood of over-indebtedness changes given the presence of the variable (with all other things remaining constant). For example, an individual with three or more children at home is likely to be nine percentage points more likely to be over-indebted than the same individual with no or fewer than three children.

Standardised estimates of the coefficients take into account the distribution (mean and variance) of the independent variables, and so are more useful when interpreting each parameter's true effect and contribution to the prediction. For simplicity, the standardised estimates have been transformed into relative importance scores that indicate the weight of each variable within the model – their absolute values sum to 100, and the sign indicates the direction of their effect.

Parameter (2018 final model)	Estimated Parameter Coefficient	Pr > Chi Sq	Standardised Estimate	Relative Importance Score	Average Marginal Probability
Intercept	*-1.517*	*<.0001*			
Private Renting	0.706	<.0001	0.085	7.6	9%
Social Renting	0.786	<.0001	0.116	10.4	10%
Aged 25-34	0.142	0.0227	0.028	2.5	2%
3+ Children	0.687	0.0225	0.029	2.6	9%
Household Income <£10k, Household Size 3+	2.262	0.0039	0.044	3.9	29%
Self-employed	2.363	0.0002	0.069	6.2	30%
Has 2+ Credit Cards	2.674	<.0001	0.135	12.1	34%
Has Savings £10,000+	-4.905	<.0001	-0.339	-30.4	-63%
Retired	-0.989	<.0001	-0.180	-16.2	-13%
Has life protection policy	-1.693	<.0001	-0.089	-8.0	-22%

Model uses 15,095 observations, of which 2,584 are over-indebted. All coefficients are statistically significant at a 95% confidence level.

The predictor factors that bear the most importance in the 2018 over-indebtedness model are those that suggest whether the individual has a significant savings balance (greater than £10,000), whether or not they are retired, and whether they have two or more credit cards.

The 2018 model is not too dissimilar from previous years. Five of its ten variables also appeared in the 2017 model and some of the new variables are very similar to previous variables. The "retired" variable is a close approximation for the previous age bands, "65-74" and "75+", and clearly the introduced age band "25-34" has a very strong (negative) correlation with these variables too.

This leaves three new variables that predict over-indebtedness, which all have a strength of reasoning. Self-employment (with its non-guaranteed income, irregular cash flow and casual work) and multiple credit cards (more monthly commitments to meet) can both be seen to influence over-indebtedness. The growth in self-employment has been much commented on in the last year and it seems appropriate

that this factor now enters the model. Life protection is also a worthy addition, as it acts as a proxy for financial stability and security – those with protection are often those on higher incomes and an increased ability to meet financial commitments.

7.3. Variable Definitions

Private Renting

The likelihood (ranging from 0 to 1) that an individual lives in a home that is rented privately.

Social Renting

The likelihood (ranging from 0 to 1) that an individual lives in a home that is rented through a local authority or housing association.

Aged 25-34

The likelihood (ranging from 0 to 1) that an individual is aged between 25 and 34 years old (inclusive). Other age bands (including broader bands) did not prove significant in any model.

Has 3+ Children

The likelihood (ranging from 0 to 1) that an individual is aged 25-39 and has three or more children at home. The inclusion of the age criteria ensures that an effect is truly caused by the presence of children and not by other age-related secondary effects. For example the very old and very young are unlikely to have more than two children at home, and so these individuals should be removed from the set with 3+ children that is being compared against. Other age criteria were examined, but 25-39 continued to provide the strongest model.

Household Income <£10k, Household Size 3+

The likelihood (ranging from 0 to 1) that an individual lives in a household of at least three people (adults or children) and that the household income is £10,000 or below.
The inclusion of the household size criteria improves the performance of this variable in two ways. Firstly, it adds an element of equivalisation, whereby larger households are treated differently to other households of similar income, under the premise that the income needs to “go further”. And secondly it helps remove other age-related secondary effects. For example, the elderly are likely to live on their own or in a couple, and are also more likely to have an income lower then £10,000. We know that older individuals are less likely to be over-indebted in general, and this interaction helps remove them from this variable and allow it to become more statistically significant.
Other income bands and household sizes were tested, but this combination produced the best model in terms of effect and significance.

Self-employed

The likelihood (ranging from 0 to 1) that an individual classifies themselves as self-employed within their main occupation, either full-time or part-time.

Has 2+ Credit Cards

The likelihood (ranging from 0 to 1) that an individual has two or more credit cards.

Has Savings £10k+

The likelihood (ranging from 0 to 1) that an individual has savings with a total value of at least £10,000. All savings products (fixed and variable) are included, but investment products and pension savings are not included.

Retired

The likelihood (ranging from 0 to 1) that an individual is not working and classifies themselves as retired. This may be with or without pension, and is not dependent on state retirement age.

Has Life Protection Policy

The likelihood (ranging from 0 to 1) that an individual has a life protection policy, including both term and whole-of-life policies

The source for each variable is given below.

Model Variable	Source of Data
Private Renting	FRS
Social Renting	FRS
Aged 25-34	FRS
3+ Children	FRS
Household Income <£10k, Household Size 3+	FRS
Self-employed	TGI
Has 2+ Credit Cards	FRS
Has Savings £10,000+	FRS
Retired	TGI
Has life protection policy	FRS

FRS = Ocean: Modelled by CACI, based on data from the *Financial Research Survey*, Ipsos
TGI = Ocean: Modelled by CACI, based on data from the *Target Group Index Research*, Kantar

8. Evaluation of Model

8.1. Statistical Significance of Parameters

As demonstrated in 0 all variables in the model are significant, at the required 95% confidence limit. In fact, the variables all demonstrate further confidence and meet a more stringent 97.5% limit.

8.2. Hosmer-Lemeshow Test

The Hosmer-Lemeshow Test is a test for goodness-of-fit within a logistic regression model. It is frequently used to evaluate predictive models of this kind by attempting to identify a "lack of fit".

The test first sorts observations (individual survey respondents) into ten equal-sized groups, based on the modelled probability of each one being over-indebted.

The expected number of over-indebted individuals within each group can be calculated by summing the modelled probabilities. These projections are then compared to the observed values in each group (counts of individuals who actually reported over-indebtedness in the research).

These ten pairs of numbers (observed versus modelled) should be close to each other, and they can be statistically tested using a Chi-square test.

Partition	Observed (Survey Data)	Modelled
1 (Respondents least likely to be over-indebted)	23	24.5
2	44	44.4
3	86	98.1
4	160	164.5
5	220	219.0
6	281	269.4
7	334	322.4
8	416	386.5
9	453	466.0
10 (Respondents most likely to be over-indebted)	567	589.2

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	Degrees of Freedom	Pr > ChiSq
7.9089	8	0.442

The test confirmed that there is no lack of fit (as the Pr>ChiSq value is larger than 0.25), and so it can be concluded that the predicted levels of over-indebtedness within the ten groups are sufficiently close to observed levels.

8.3. *C*-Statistic

The *c*-statistic, or "concordance statistic" is a common test to report on within logistic regression analysis, and is a single measure of the reliability of the predicted levels of over-indebtedness, *at an individual level*.

As the objective of this model is to provide expected levels of over-indebtedness at a local area level (by summing individual-level probabilities), individual-level predictions are less relevant, so this test is less relevant than the Hosmer-Lemeshow goodness-of-fit test. However if a model can be demonstrated to have good concordance – i.e. where those that are over-indebted receive higher likelihoods than those that aren't – then this gives us additional confidence that our aggregated area estimates are more likely to be accurate.

Each "over-indebted" observation (i.e. survey respondents who said they were over-indebted) is paired with every "not over-indebted" observation. In the modelled data set of 15,095 usable observations, the observed number of over-indebted individuals is 2,584 and the number of non-over indebted individuals is 12,511. This generates 32,328,424 (12,511 x 2,584) possible pairings of an over-indebted individual with a not over-indebted individual. In each pairing, the predicted likelihoods of being over-indebted can be compared. If the model provided a reliable prediction, then the likelihood for the over-indebted individual should always be greater than the likelihood for the not over-indebted individual (this is known as "concordance"). And if the model is entirely random, it would be expected for this to only occur in half of the pairings.

Percent Concordant	73.4%
Percent Discordant	26.2%
Percent Tied	0.4%
Pairs	32,328,424

Somers' D	0.471
Gamma	0.473
Tau-a	0.134
c	0.736

The *c*-statistic for the over-indebtedness model is 73.6%.

In other words, if an over-indebted (A) and a not over-indebted individual (B) were randomly selected from the survey respondents, the model is likely to give (A) a higher likelihood of being over-indebted than (B). If this was done 100 times, the model would correctly give the over-indebted individual a higher probability on 73 (or 74) occasions.

A model is considered good if $c > 70\%$ and strong when it is $> 80\%$ (Hosmer & Lemeshow, 2013). This is an acceptable result for the modelling objectives as previously explained, and passes the model acceptance criteria of 70%. Furthermore, it is an improvement on the 2017 model's concordance of 70.9%.

8.4. Multicollinearity

The variables selected in the model should be statistically independent. In other words, there should be no strong correlation between any pairs of variables. This can be tested by creating a correlation matrix of the variables. The score (Pearson's correlation moment) ranges from -1 to 1. A score of -1 indicates a

perfect negative correlation, 1 indicates a perfect positive correlation, and scores close to 0 indicate no correlation at all. Strong correlations are normally indicated by scores greater than 0.7 (or less than -0.7), but it is prudent to examine moderate correlations too – those with scores larger than 0.5 (or smaller than -0.5).

Model Variable		1	2	3	4	5	6	7	8	9	10
1	Private Renting	1.00	0.46	-0.04	0.22	-0.06	-0.12	-0.44	**0.51**	-0.24	**0.51**
2	Social Renting		1.00	-0.01	0.04	-0.40	-0.25	-0.10	0.35	-0.31	0.16
3	Aged 25-34			1.00	-0.08	0.13	0.17	-0.23	0.19	-0.02	0.35
4	3+ Children				1.00	-0.11	-0.06	-0.17	0.27	0.00	0.02
5	Household Income <£10k, Household Size 3+					1.00	0.16	0.19	-0.07	0.23	0.22
6	Self-employed						1.00	-0.25	-0.11	0.33	-0.05
7	Has 2+ Credit Cards							1.00	**-0.68**	0.45	-0.37
8	Has Savings £10,000+								1.00	**-0.58**	0.30
9	Retired									1.00	-0.19
10	Has life protection policy										1.00

Some moderate multicollinearity is to be expected in logistic regression models, however the model presents only four incidences greater than 0.5, and worthy of further attention.

The strongest correlation (-0.68) is between individuals who have two or more credit cards and those that (don't) have savings of at least £10,000. Although this score suggests a reasonably strong correlation, both variables are strongly significant (at a confidence level greater that 99.99%), with strong effects that act in opposite directions. Coupled with the negative correlation, this is not likely to introduce over-fitting into the model.

There is some relationship between those without savings of £10,000 and those who are retired. However the correlation score is just -0.58, indicating only a moderate relationship. The remaining two incidences are private renting with both the savings variable and life protection variable – both correlations are 0.51 and offer no cause for concern.

8.5. Model Validation

The research data was split into two parts: an 80% "training" sample to be used to create a revised over-indebtedness model, and a 20% "validation" sample that can be applied to the model to independently verify the accuracy and suitability. The data was sampled using a stratified 1-in-5 selection method ensuring demographic and regional representativeness.

Of 3,774 records in the validation sample, 634 were classified as over-indebted (16.8%). The 2018 model predicts an over-indebtedness figure of 650 individuals (17.2%). This represents a 2.5% over-prediction, or a 0.42 percentage point error in the over-indebtedness rate – well within the acceptance criteria (within 10% of the reported level).

This is an improvement on the 2017 result, where the modelled error on the validation sample was 0.7 percentage points.

Model Variable	Count	%
Usable Records	3,774	
Modelled Over-indebtedness:	650	17.2%
Observed Over-indebtedness:	634	16.8%
Modelled Error (%)	2.50%	
Modelled Error (pp)	0.42%	

8.6. Stability of Model and Prediction

2018 represents the first real development of the model in term of input variables. A fuller revision is typical after three to four years in models of this type, and with a constantly changing economy and financial context it is right that we periodically investigate new factors of over-indebtedness.

9. Contacts

CACI

Technical Author	Jamie Morawiec	020 7605 6035	jmorawiec@caci.co.uk
Account Manager	Henry Steenstra	020 7605 6201	hsteenstra@caci.co.uk

The Money and Pensions Service

Insight Manager	Sarah Little	020 8132 4954	sarah.little@maps.org.uk