the Money Advice Service

A Picture of Over-Indebtedness

Technical report

Contents

Background and Objectives

The Money Advice Service has been measuring individuals' levels of over-indebtedness as part of research surveys since 2012. As part of the Money Advice Service's business plan, there is a requirement to generate a geographic measure of over-indebtedness that can be applied to every adult in the United Kingdom. The output would assist the Money Advice Service in a number of areas:

- Forecasting service demand;
- Mapping and managing funding/resource; and
- Understanding factors associated with over-indebtedness.

The Money Advice Service worked with CACI to produce a nationwide model, which combined large numbers of survey respondents with CACI's rich consumer data, to produce over-indebtedness scores for a range of geographies. The approach was a "bottom-up" methodology, meaning individuals were modelled separately and then aggregated into regions based on their residential postcode. This granular method is not only more robust than modelling "local averages", but is also more flexible and allows other geographies (or indeed individual or household-level) analysis in the future.

The brief required the solution to be clear and understandable, taking a transparent approach to the way it is calculated and delivering robust local data. The following report details the approach and output of the Money Advice Service over-indebtedness model.

Data Sources

Research Data

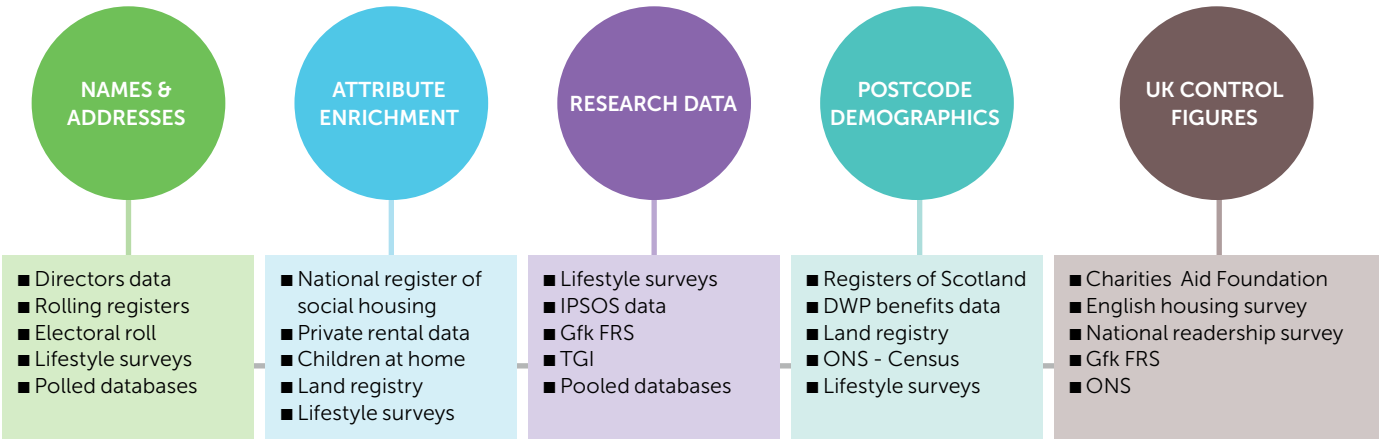
The Money Advice Service provided CACI with research survey data for analysis. The total sample size across all sources was approximately 16,000 individual respondents.

The majority of the research data came from YouGov's quarterly "Debt Tracker" survey, and this was topped-up with two omnibus surveys. The remainder came from the Money Advice Service's 2015 Financial Capability Strategy (FinCap) survey. The time period covered across all research was November 2014 to August 2015, and survey respondents were de-duped across surveys and waves.

Survey	Conducted by	Sample Size	Mode	Period
Debt Tracker	YouGov	6,138	Online	November 2014 / February 2015 / May 2015 / August 2015
Omnibus	YouGov	4,276	Online	15th June 2015 / 16th June 2015
FinCap	GfK	5,555	74% Online, 26% Face-to-face	April 2015

Ocean Data

Ocean is an attribute-rich consumer database for the UK, updated quarterly. Hundreds of millions of records from research surveys, open data, government data and many other sources are collated together to create Ocean.



Ocean includes:

- Names and addresses of 51 million adults. 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 different high-volume sources, and selecting the most up to date information.
- 29 million dates of birth. Dates of birth are selected and combined from multiple sources. Where date of birth isn't available, age is imputed from a range of data including forename, household composition, region, product holdings and other lifestyle variables.
- A wide range of variables for each individual. Variables can be supplied as hard facts (i.e. indicator variables) where known or where they can be confidently imputed from other knowledge about the individuals. Whenever they are not known for certain, values are inferred from modelling based on other known characteristics. Modelled estimates can be provided as estimates of the probability that a person has an attribute, as inferred Yes/No flags, or as categorical assignments for appropriate variables such as tenure.
- A set of composite indicators built by combining a number of variables to indicate attributes such as wealth, disposable income, etc.

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

ATTRIBUTES

- Age and sex
- Number and age of children
- Income
- Household size and composition
- Length of residence
- Housing: type, tenure, size, value
- Occupation
- Social Grade
- Number, age and type of cars
- Use of glasses and contact lenses

ATTITUDES

- Reading preferences; books and magazines
- Charities: which causes supported and how
- Newspaper readership
- Attitudes to financial products and channels
- Intention to switch financial products
- Attitudes to online privacy and safety
- Lifestyle attitudes
- Shopping attitudes

BEHAVIOUR

- Technology ownership and use
- Holidays: destination, type, spend and booking method
- Interests and hobbies
- Supermarkets: weekly spend and brands
- Mail order: frequency and kind of goods bought
- Financial products owned
- Savings and Investments value
- Credit card patterns of use
- Medical insurance
- Smokers
- Internet usage: frequency, location and technology
- Types of goods and services researched online
- Types of goods and services purchased online
- Online activities: gambling, dating, gaming etc.
- Social networking: which networks and types of activity
- Music / video downloads
- Mobile phone: type of phone and how used
- Shopping: types of stores visited (premium, mass, value)
- Environmentally friendly behaviour

Defining "Over-Indebtedness"

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

The definition of "over-indebted" was derived from the same two questions within the FinCap and YouGov surveys. Respondents must have answered either:

i. I find meeting my monthly bills/commitments a ***heavy burden;*** and/or

ii. I have missed bill 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.[1] Those that do not respond positively to either question were defined as "not over-indebted".

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

Within the supplied surveys (which were weighted to be representative of the UK), the average proportion of respondents finding bills a heavy burden was 11.0%, while 7.6% of respondents had been in arrears in three of the last six months.

The overlap between respondents who answered yes to both questions was 3.6%, thus the overall observed level of over-indebtedness in the surveys was 15.0%.

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

Survey Source	Respondents	Paying bills is a burden	Arrears in 3m of last 6m	Over-Indebted
YouGov / Debt Tracker	6,138	10.13%	5.29%	12.56%
YouGov / omnibus	4,276	14.15%	7.27%	16.84%
FinCap	5,555	9.65%	10.34%	16.21%
Total	**15,969**	**11.04%**	**7.58%**	**14.97%**

1. "To what extent 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]. Question source: Debt Tracker q60 / Omnibus q1 / FinCap c1

"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 months." [Yes; No ; Don't know] . Question source: Debt Tracker q460 / Omnibus q2 / FinCap c2

Tagging Respondents with Ocean Data

The first stage of the model-build was to match the 16,000 survey respondents to CACI's database of individuals. This appended the full range of Ocean attributes and characteristics to each respondent, so that variables could be tested against the dependent over-indebtedness variable.

There are three levels of match:

- Individual or Household: respondent is matched to a known individual, or a known household, within the Ocean database.
- Postcode: Respondent cannot be matched to a known individual or household with Ocean (normally because of an absence of name/address in the survey). Where a valid postcode is supplied, some characteristics can be imputed.
- Unmatched: Respondent cannot be matched to a known individual or household within Ocean, and no valid postcode has been provided.

The levels and quality of match across the three survey sources varied, but overall provided a good match-rate. A large proportion of respondents (78%) were able to be matched to Individual or Household, meaning there is confidence in the Ocean attributes assigned to these individuals. Respondents matched only by postcode were not taken forward for modelling, because of the lower accuracies of the characteristics attributed to them. After de-duping, approximately 11,000 respondents matched at individual-level were then carried forward to the modelling process.

Survey Source	Individual or Household Match	Postcode Match	Unmatched
YouGov / Debt Tracker	79%	17%	4%
YouGov / Omnibus	73%	25%	1%
Fincap	80%	18%	2%
Total Proportion	**78%**	**19%**	**3%**

The total sample used for the model was 11,279 which is a result of de-duping the total number of records matched at Individual or Household level.

Modelling Process

Model Suitability

Dependent Variable

The dependent variable (that which is being modelled) is the event that an individual is over-indebted. Thus, the resultant binary variable "INDEBTEDNESS" gives 1 where an individual is over-indebted and 0 otherwise. Logistic regression is a proven statistical method for analysing binary data.[2]

Adequate Sample Size

The matched and de-duped sample size is 11,279. Using the often-quoted rule of thumb that a model requires fifty observations for every independent variable (Peduzzi et al, 1996), this is a sufficient size to avoid over-fitting.

This sample size is therefore adequate for a robust model.

Variables Tested

CACI tested all Ocean variables, including a wide variety of aggregations and interactions, against the over-indebtedness variable. In addition Fresco and Acorn (demographic segmentation products) were also tested in the full model.

The first stage of the modelling process was to look at the correlations between over-indebtedness and each Ocean variable in isolation. This gave an early indication of likely predictors, and also provided validation to variables in the final model by way of univariate analysis.

The below gives a summary of some of the variables tested within the model, and the most over and under-indexed Ocean variables are given in appendix 11.1.

- Regional indicator variables and urban/rural flags
- Age (in 5-year bands) and gender
- Household composition (e.g. single person, married family) and household size
- Number and age (5-year bands) of children at home
- House type (e.g. bungalow), house age (20-year bands), number of bedrooms
- House value (seven value bands)
- Social grade (A-E), employment status, and occupation type
- Level of education
- Household tenure (e.g. own with mortgage, rent privately)
- Household income (£10k bands)
- Car ownership: number, age and type
- Financial product holdings (e.g. credit card, loan for consolidation)
- Value of savings and value of investments (both within ten value bands)
- Financial attitudes (e.g. regularly read financial pages, trust price comparison sites)

2. All statistical modelling was carried out using SAS 9.3, operating in a Windows environment.

Many other lifestyle (eg holiday and interests), buying (eg supermarket and newspaper readership), channel preference, technology & social media, and health variables were also tested.

In total, almost 500 individual single variables were considered within the analysis.

Modelling Approach

A stepwise variable selection method was used to generate an initial model, which provided an early indication of potential driving factors, as well as a likely quality of model that would be possible. The stepwise model was carried out with a 0.1 alpha for entry and a 0.15 alpha for exit. This means that variables were added to the model if they were significant at the 90% level, and were removed if their level fell to 85% or lower at a later iteration.

Due to the limited capabilities of a stepwise algorithm in relation to interactions and classing variables, manual testing was required to build on and improve the over-indebtedness model.

All variables showing a strong (positive or negative) relationship with over-indebtedness amongst the survey respondents were tested in the model. In addition they were also tested for regional interactions. Variables (and their interactions where appropriate) were retained within the model where they contributed positively to the model (improved the model fit statistics) and were significant at the 90% confidence level.

Examples of interactions tested include the below. These were not included in the final model as they provided no improvement to fit statistics.

- Being unemployed and owning a home with a mortgage
- Aged under 25 and having at least one child
- Private renting in London

Model Parameters

The recommended model consists of sixteen independent variables, some of which are combinations and interactions of individual Ocean variables. Twelve are positive factors (suggest an increased likelihood of over-indebtedness), and four are negative factors (suggest a decreased likelihood of over-indebtedness).

The sign of the parameter coefficients indicates whether the variable has a positive or negative effect on over-indebtedness. To understand the magnitude of a change in the dependent variable to the likelihood of being over-indebted, 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 a loan for consolidation is likely to be 51 percentage points more likely to be over-indebted than the same individual without a loan.

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 one, and the sign indicates the direction of their effect.

Parameter	Estimated Parameter Coefficient	Standardised Estimate	Relative Importance Score	Average Marginal Probability
Intercept	-1.925			
Has loan for consolidation	4.584	0.047	4.2	51%
Private renting	0.315	0.045	4.0	3%
Social renting	0.431	0.074	6.6	5%
Has 3+ children	1.050	0.029	2.6	12%
Single parent	0.209	0.027	2.4	2%
Social Grade D or E	1.067	0.093	8.3	12%
Northern Ireland	0.527	0.063	5.6	6%
Value of home <£100k, South East	0.831	0.035	3.1	9%
Value of home <£100k, London	4.464	0.022	1.9	49%
Unemployed, Wales & West Midlands	1.952	0.042	3.8	22%
Own home outright, Wales	0.670	0.054	4.7	7%
Household income <£10k, Household size 3+	1.159	0.038	3.4	13%
Has savings £10k+	-2.127	-0.200	-17.8	-24%
Aged 65-74	-0.919	-0.185	-16.4	-10%
Aged 75+	-1.211	-0.126	-11.2	-13%
Scotland	-0.259	-0.046	-4.1	-3%

Model uses 11,279 observations, of which 1,551 are over-indebted

Two variables were examined more closely to assess their contribution to the model. Both "Loan for Consolidation" and "Value of Home <£100K, London" have very high marginal effects, however the overall number of individuals displaying these characteristics is small (2% and 1% respectively). In the few cases these characteristics are present, the effect is very strong, which is showed by the average marginal probability of fifty percentage points. Additionally the statistical significance of loan for consolidation is very strong (p-value = 0.005, see section 7.2).

Due to the presence of the "Value of Home <£100k, South East" variable, it was felt that the equivalent variable for London was required to provide geographical continuity (since the South East wraps around London). However it was important to keep these two regions separate, since the effect on over-indebtedness is much stronger in London than the South East, as seen in the parameter coefficients and marginal probabilities.

Models without the variables "has loan for consolidation" and "Value of Home < £100k, London" were also analysed, but showed weaker fit statistics – increased Akaike Information Criterion, smaller concordance, and weaker goodness-of-fit (Hosmer-Lemeshow) statistic. It was therefore decided that these variables were retained in the model.

Variable Definitions

Has Loan for Consolidation

The likelihood (ranging from 0 to 1) that an individual has an unsecured loan for the purpose of consolidating existing debt.

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.

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 also examined, but 25-39 provided the strongest model.

Single Parent

The likelihood (ranging from 0 to 1) that an individual lives in a single-person household (defined as one adult present), and that there is at least one child (aged 0-17) also living at the address.

Northern Ireland

An indicator variable (0 or 1) depending on whether the individual's residential postcode falls within the government region of Northern Ireland.

This variable is an adjustment factor to account for the relatively higher levels of over-indebtedness in Northern Ireland.

Social Grade D or E

The likelihood (ranging from 0 to 1) that an individual is classified within the NRS social grades D or E.

Value of Home <£100k, South East

The likelihood (ranging from 0 to 1) that the value of an individual's home is less than £100,000 and that they live in the South East of England (excluding Greater London).

Value of Home <£100k, London

The likelihood (ranging from 0 to 1) that the value of an individual's home is less than £100,000 and that they live in Greater London.

Unemployed, Wales & West Midlands

The likelihood (ranging from 0 to 1) that an individual is unemployed and that they live in either Wales or the West Midlands.

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.

Other income bands and household sizes were tested, but this combination produced the best model in terms of effect and significance.

Own Home Outright, Wales

The likelihood (ranging from 0 to 1) that an individual owns their home outright (i.e. without a mortgage) and that they live in Wales.

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 are not included.

Investment values and individual product holdings were also tested, but these introduced multicollinearity into the model and compromised overall fit.

Aged 65-74

The likelihood (ranging from 0 to 1) that an individual is aged between 65 and 74 years old (inclusive).

Other age bands (including broader bands) did not prove significant in any model.

Aged 75+

The likelihood (ranging from 0 to 1) that an individual is aged 75 years or older.

Other age bands (including broader bands) did not prove significant in any model

Scotland

An indicator variable (0 or 1) depending on whether the individual's residential postcode falls within the government region of Scotland.

This variable is an adjustment factor to account for the relatively lower levels of over-indebtedness in Scotland.

The above variables were all tested for interactions with each other, and with each of the regional variables. Those which provided better fit statistics retained their interaction, whilst those that didn't (or else produced similar parameter coefficients across regions) had the interactions removed.

The source for each variable is given in appendix 11.2.

Evaluation of Model

Global Null Hypothesis Test

This test examines the hypothesis that all of the parameter coefficients are zero, and that the variables in the model do not determine the level of over-indebtedness at all.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood ratio	762.7105	16	<.0001
Score	719.5014	16	<.0001
Wald	605.1717	16	<.0001

The test (using three different measures) showed that the parameters are significantly different from zero, indicating that the model is a valid prediction of over-indebtedness.

Statistical Significance of Parameters

All variables in the model are significant, using a 90% confidence limit, which is an acceptable level when analysing collated survey data. In fact, the confidence can be increased for the majority of the variables – up to 99% in some cases. The table below shows confidences for individual parameters.

Parameter	Wald-Chi Square	Pr>ChiSq
Intercept	273.540	<.0001
Has loan for consolidation	7.909	0.005
Private renting	7.881	0.005
Social renting	9.760	0.002
Has 3+ children	4.655	0.031
Single parent	3.789	0.052
Social Grade D or E	10.257	0.001
Northern Ireland	19.925	<.0001
Value of home <£100k, South East	4.455	0.035
Value of home <£100k, London	2.467	0.096
Unemployed, Wales & West Midlands	5.688	0.017
Own home outright, Wales	7.991	0.005
Household income <£10k, Household Size 3+	4.896	0.027
Has savings £10k+	40.697	<.0001
Aged 65-74	41.512	<.0001
Aged 75+	19.320	<.0001
Scotland	7.183	0.007

There are two variables included that have a p-value larger than 0.05. "Value of home <£100k, London" and "single parent". On one hand, "value of home <£100k, London" occurs in very small incidences, but with a very large marginal effect (50 percentage points). It also works alongside the South East variable to produce a contiguous geographical region for which low home value applies. Besides as explained in section 5, excluding this variable generates a weaker fit statistics. On the other hand, the p-value of "single parent" is borderline at a 95% confidence level, but models omitting this variable performed worse in predicting over-indebtedness. Therefore, we conclude to include them in the model.

Hosmer-Lemeshow Test

The Hosmer-Lemeshow Test assess 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 equal-sized groups, based on the modelled probability of each one being over-indebted. The number of groups is defined by the number of covariant terms (independent variables) plus one, which in this case is 17 groups. Ten groups (deciles) were also tested (the standard in most statistical packages) and also confirmed no lack of fit – these results are available on request.

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 said they were over-indebted in the research surveys).

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

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
10.7027	15	0.7734

The test confirmed that there is no lack of fit (a Pr>ChiSq value larger than 0.1), and so it can be concluded that the predicted levels of over-indebtedness within the groups are sufficiently close to observed levels.

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*. However, because 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. The Hosmer-Lemeshow goodness-of-fit test is a much more appropriate test in this case, but the concordance statistic has been included in this report for completeness.

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 11,279 usable observations, the observed number of over-indebted individuals is 1,551.

Over-indebted	Respondents
0	9,728
1	1,551
Total	**11,279**

This generates 15,088,128 (9,728 x 1,551) 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.

	Number of Pairs	% of Pairs
Pairs Concordant	10,722,921	71.1%
Pairs Discordant	4,365,161	28.9%
Pairs Tied	46	0.0003%
Total Pairs	**15,088,128**	

The c-statistic for the over-indebtedness model is 71.1%.

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 71 occasions.

A model is considered good if c > 70% and strong when it is >80% (Homer & Lemeshow, 2000). This is an acceptable result for the modelling objectives as previously explained.

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.

	Model Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Has loan for Consolidation	1.00	0.01	0.00	0.21	0.15	-0.06	-0.03	0.06	-0.01	-0.04	-0.07	-0.04	-0.28	-0.35	-0.21	-0.04
2	Private renting		1.00	-0.16	0.10	0.16	0.01	0.05	0.04	0.02	0.04	0.13	-0.08	-0.29	-0.17	-0.09	-0.05
3	Social renting			1.00	0.06	0.09	-0.04	0.75	0.06	0.06	0.06	0.37	-0.08	-0.34	0.04	0.01	-0.01
4	Has 3+ children				1.00	0.17	0.02	0.02	0.00	0.01	0.02	0.12	-0.05	-0.25	-0.16	0.07	-0.01
5	Single parent					1.00	0.03	0.12	0.01	-0.01	0.02	0.18	-0.03	-0.19	-0.15	-0.07	-0.02
6	Northern Ireland						1.00	-0.03	-0.02	-0.01	-0.05	0.02	-0.05	0.00	-0.03	-0.02	-0.08
7	Social Grade D or E							1.00	0.09	0.06	0.13	0.45	-0.01	-0.53	-0.01	0.02	0.04
8	Value of home <£100k, South East								1.00	0.00	-0.02	0.02	-0.02	-0.04	0.00	-0.01	-0.03
9	Value of home <£100k, London									1.00	-0.01	0.03	-0.01	-0.03	0.00	-0.01	-0.02
10	Unemployed, Wales & West Midlands										1.00	0.12	0.19	-0.10	-0.01	0.00	-0.07
11	Household income <£10k, Household Size 3+											1.00	-0.03	-0.37	-0.16	-0.08	-0.02
12	Own home outright, Wales												1.00	0.06	0.08	0.03	-0.07
13	Has savings £10k+													1.00	0.51	0.16	-0.08
14	Aged 65-74														1.00	-0.08	-0.02
15	Aged 75+															1.00	-0.02
16	Scotland																1.00

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

The strongest correlation (0.75) is between individuals who rent their home socially and those who are classed in social groups D and E. Although this score suggests reasonably strong correlation, both variables are strongly significant, with strong positive effects.

The second-strongest correlation is the negative relationship between social grade and high savings values, with a correlation of -0.53. This in principle should not affect the model, but more importantly social grade E includes retired people – a demographic very likely to have high value savings. We must include both variables in a model to ensure that we can distinguish pensioners that have no/few savings, in comparison to those with high values of savings.

Additionally, standard errors[3] for these variables showed no signs of abnormality, which would be the case if multicollinearity was affecting the model estimates of over-indebtedness.

Cross-Validation

The model was cross-validated by employing the "leave one out" principle. This consists of removing a single observation one at a time, and re-estimating each of the model parameters. These revised model-parameters yielded a probability of over-indebtedness sufficiently close to the original value for each observation. Therefore it can be concluded that the model is not over fitted, and would perform well on a generalised population.

It was not considered appropriate to validate the model using a blind testing sample. This process involves randomly splitting the survey respondents into two separate data sets (ideally 70% and 30% partitions). The model is defined on the larger partition, and then tested on the second smaller partition. However with 11,000 survey respondents to begin with, it was felt the quality of the model would be compromised by reducing the number of observations. Instead the next available wave of the YouGov Debt Tracker survey was used to validate the model (see section 7.7).

Out of Sample Validation

The model was validated against a separate wave of survey respondents. This was the November 2015 wave of YouGov's Debt Tracker, the next successive wave following those which were used in the model.

After de-duping the respondents against those within the model build, there were 1,976 unique and previously unseen individuals. 76% of these were coded and accurately matched to CACI's Ocean database, leaving an out of sample validation set of 1,496 individuals.

	In sample data	Out of sample validation data
Sources	YouGov Debt Tracker, YouGov Omnibus, FinCap	YouGov Debt Tracker
Time period	Nov 2014 – Aug 2015	Nov 2015
Records	15,969	1,979
Usable records	11,279	1,496
Observed over-indebted	15.3%	12.4%

3. Standard errors for model variables are available on request

The observed over-indebtedness amongst the out of sample validation set was 12.4%. This is significantly lower than that of the original in sample data set used to construct the model (15.0%). It was shown that the respondents in the out of sample validation set were both regionally and demographically representative of the training set as a whole.

The sixteen-variable model was applied to the out of sample data, and this yielded an overall over-indebtedness estimate of 200 adults, 13.4%.

Modelled Over-indebtedness within out of sample set:	200 adults	13.37%
Observed Over-indebtedness within out of sample set:	186 adults	12.43%

Given the relatively small sample size of 1,496, an error of just 0.94 percentage points on the over-indebtedness level can be interpreted as a good performance of the model. This is especially the case given the low incidence of over-indebtedness (186 adults), where the model was able to predict a figure of 200 – an error of just fourteen adults (or 7.5%).

Stability of Prediction (Alternative Models)

Amongst the alternative models tested (e.g. those without the variables "loan for consolidation" and "value of home <£10k, London"), the various levels of over-indebtedness forecast had a range of no more than 0.2%. This is further validation of the robustness of the model, and the insensitivity of the prediction to certain variables.

Model	UK Forecast of Over-indebtedness	AIC Score
BASE MODEL	**16.1%**	**8,304**
Omitting "value of home <£100k, London"	16.1%	8,304
Omitting "loan for consolidation"	16.3%	8,309
Including variables for all regions	16.2%	8,312
Including variable for all age bands	16.1%	8,306

In the table above, the "base model" is the final over-indebtedness model containing the sixteen variables.

The next two models each remove one variable, the inclusions of which are discussed in section 6. Removing "value of home <£100k, London" had no effect on the over-indebtedness prediction, whilst removing "loan for consolidation" increased it marginally. The final two models include dummy variables into the model for each of the twelve regions and seven age bands. This was to test whether the base model was missing any effect from either age or region. In both cases the overall level of over-indebtedness predicted was stable compare to the more parsimonious base model.

In all cases the model fit statistics provided no improvement on the base model (as indicated by the increased AIC scores in the table), and therefore these alternative models were rejected.

Resulting Output

Applying the Model to the UK Population

The over-indebtedness model was built on 11,279 survey respondents from across the UK. Because the independent variables are all available within CACI's Ocean database, the model could be applied to the 51 million adults in the UK at an individual level. For the purpose of these counts, "adults" are defined to be individuals aged 18 years or older.

Over-indebtedness scores (the likelihood of being over-indebted) were first built at individual level from the count of Ocean adults. These were then applied to the latest 2015 population estimates (at unit postcode level) to produce a definitive projection of over-indebted individuals for each postcode. Where required, over-indebted counts were adjusted to the latest and most accurate population estimates at a unit postcode level, before being aggregated to areas.

CACI are the sole data provider to the Joint Industry Committee for Population Standards (JICPOPS), which ensures comparable population statistics across the advertising and media industry. These population estimates are very much seen as the standard across a wide range of industries, and are considered the most robust current year estimates available. The over-indebtedness scores have been applied to these figures to ensure the resultant area statistics are as up-to-date and accurate as possible.

UK Over-Indebtedness Figure

The headline figure for the number of over-indebted adults in the United Kingdom is 8.25 million.

This equates to 16.1% of the adult population who are regularly missing monthly payments or finding meeting commitments a heavy burden.

	Adults	Over-Indebted %	Over-Indebted Adults
United Kingdom	51,134,629	16.14%	8,254,162

Over-Indebtedness by UK Region

There is a range of over-indebtedness across the twelve regions of the United Kingdom, with average levels of over-indebtedness ranging from 13.2% in Scotland to 21.0% in Northern Ireland.

Region	Adults	Over-Indebted %	Over-Indebted Adults
Northern Ireland	1,413,461	20.97%	296,463
Wales	2,471,479	19.56%	483,330
West Midlands	4,470,400	17.98%	803,767
North East	2,096,910	17.66%	370,415
London	6,693,252	17.38%	1,163,618
Yorkshire and The Humber	4,241,681	17.14%	727,001
North West	5,630,616	16.89%	950,815
East Midlands	3,681,441	16.16%	594,952
South West	4,362,296	14.49%	632,281
East of England	4,750,639	14.38%	683,086
South East	6,999,559	13.97%	977,902
Scotland	4,322,895	13.20%	570,530

Other Geographies

Over-indebtedness has been aggregated from postcode level into three small-area geographies.

Lower Tier Local Authority

Non-metropolitan districts, metropolitan boroughs, London boroughs and unitary authorities of England. Includes all districts of Scotland (32), Wales (22) and Northern Ireland (11).

There are 391 lower tier local authorities across the United Kingdom.

The 10 most over-indebted and 10 least over-indebted local authorities (lower level) are:

Rank	Local Authority	Over-Indebted %
1	Sandwell	24.67%
2	Blaenau Gwent	24.31%
3	Merthyr Tydfil	24.13%
4	Newham	23.83%
5	Derry and Strabane	23.76%
6	Barking and Dagenham	23.02%
7	Belfast	22.94%
8	Tower Hamlets	22.94%
9	Kingston upon Hull, City of	21.94%
10	Rhondda Cynon Taf	21.89%

Rank	Local Authority	Over-Indebted %
382	Hart	10.75%
383	Aberdeenshire	10.66%
384	Epsom and Ewell	10.64%
385	Chiltern	10.61%
386	South Bucks	10.58%
387	Mole Valley	10.31%
388	Elmbridge	10.20%
389	East Dunbartonshire	10.11%
390	East Dorset	10.10%
391	East Renfrewshire	10.04%

Upper Tier Local Authority

Non-metropolitan counties, metropolitan counties, Inner & Outer London and unitary authorities.

Includes all districts of Scotland (32), Wales (22) and Northern Ireland (11).

There are 156 upper tier local authorities across the United Kingdom.

The 10 most over-indebted and 10 least over-indebted local authorities (upper level) are:

Rank	Local Authority	Over-Indebted %
1	Blaenau Gwent	24.31%
2	Merthyr Tydfil	24.13%
3	Derry and Strabane	23.76%
4	Belfast	22.94%
5	Kingston upon Hull, City of	21.94%
6	Rhondda Cynon Taf	21.89%
7	Nottingham	21.78%
8	Mid Ulster	21.72%
9	Stoke-on-Trent	21.71%
10	Caerphilly	21.49%

Rank	Local Authority	Over-Indebted %
147	Surrey	11.54%
148	Edinburgh, City of	11.54%
149	Orkney Islands	11.36%
150	Windsor and Maidenhead	11.25%
151	Perth & Kinross	11.25%
152	Eilean Siar	11.22%
153	Wokingham	10.82%
154	Aberdeenshire	10.66%
155	East Dunbartonshire	10.11%
156	East Renfrewshire	10.04%

Parliamentary Constituency

2013 Westminster parliamentary constituency boundaries, derived for the 2015 general election.

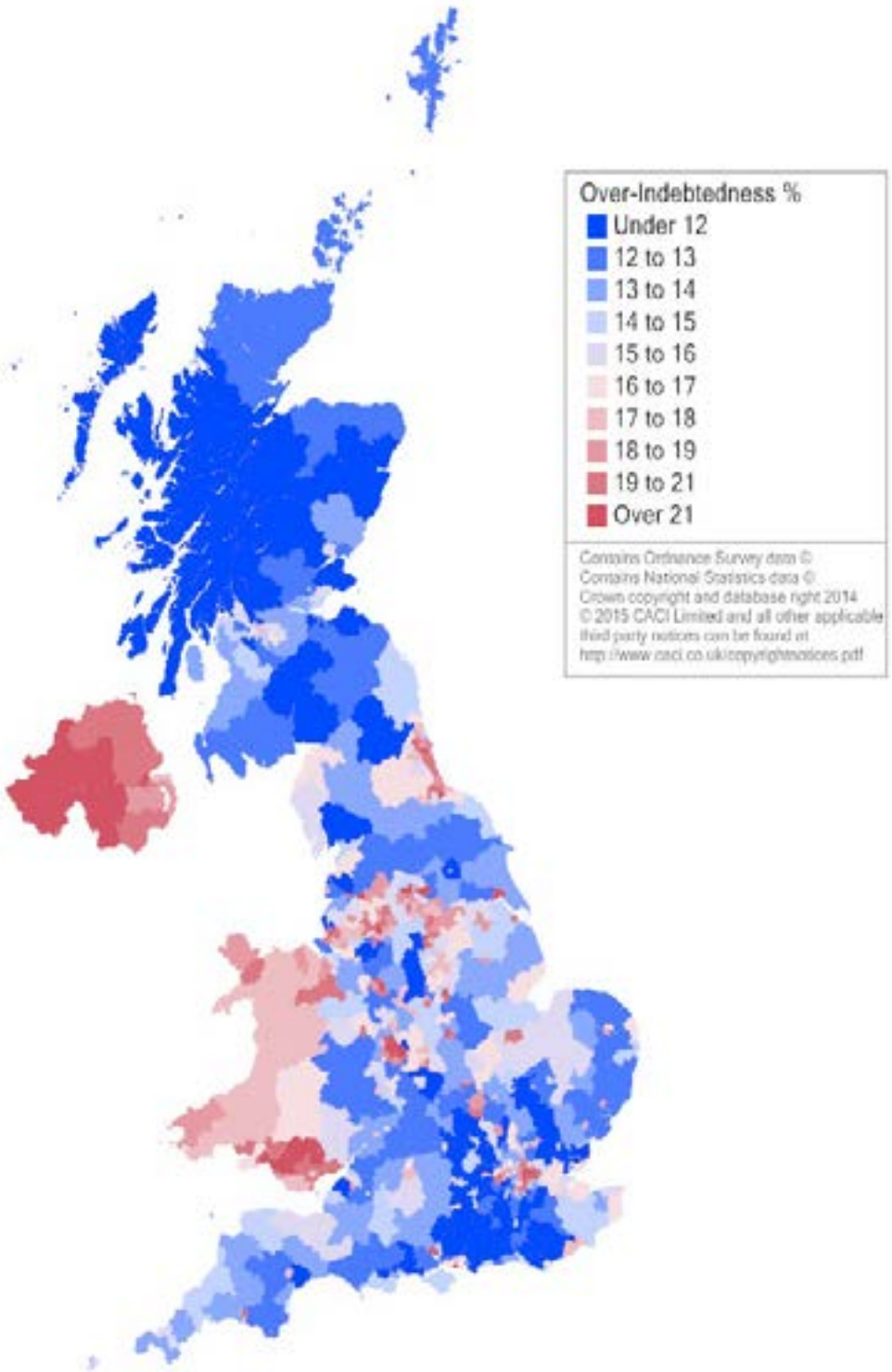
There are 650 parliamentary constituencies across the United Kingdom.

The 10 most over-indebted and 10 least over-indebted constituencies are:

Rank	Constituency	Region	Over-Indebted %
1	Birmingham, Ladywood	West Midlands	28.06%
2	Belfast West	Northern Ireland	26.80%
3	West Bromwich West	West Midlands	25.61%
4	Warley	West Midlands	24.86%
5	Birmingham, Hodge Hill	West Midlands	24.73%
6	West Ham	London	24.60%
7	Merthyr Tydfil and Rhymney	Wales	24.50%
8	Blaenau Gwent	Wales	24.31%
9	Foyle	Northern Ireland	24.18%
10	West Bromwich East	West Midlands	24.03%

Rank	Constituency	Region	Over-Indebted %
641	Esher and Walton	South East	10.29%
642	Epsom and Ewell	South East	10.27%
643	Edinburgh West	Scotland	10.19%
644	Arundel and South Downs	South East	10.18%
645	Mole Valley	South East	10.13%
646	East Renfrewshire	Scotland	10.04%
647	Edinburgh South	Scotland	9.97%
648	Gordon	Scotland	9.95%
649	West Aberdeenshire and Kincardine	Scotland	9.48%
650	East Dunbartonshire	Scotland	9.08%

Parliamentary constituencies, mapped by over-indebtedness



References

- SAS Institute Inc. *SAS/Stat User's Guide (2015)*. Available at: http://support.sas.com/documentation/93/index.html
- Delwiche Lora D and Susan J. Slaughter (2003). *The Little SAS Book: A Primer, Third Edition*, Cary, NC: SAS Institute Inc. ISBN *978*-1-59047-333-7
- Hosmer, David W.; Lemeshow, Stanley (2013). *Applied Logistic Regression*. New York: Wiley. ISBN 978-0-470-58247-3
- Alan Agresti (2012). *Categorical Data Analysis*. Hoboken: John Wiley and Sons. ISBN *978-0-470-46363-5*
- Peduzzi, P; Concato, J; Kemper, E; Holford, TR; Feinstein, AR (1996). *A simulation study of the number of events per variable in logistic regression analysis*. Journal of Clinical Epidemiology 49 (12)

Contacts

CACI

Technical Author:
Jamie Morawiec
020 7605 6035
jmorawiec@caci.co.uk

Analysis Director:
Richard Tomlinson
020 7605 6171
rtomlinson@caci.co.uk

Account Manager:
Henry Steenstra
020 7605 6201
hsteenstra@caci.co.uk

The Money Advice Service

Strategy and Innovation Executive:
Jair Munoz-Bugarin
020 7943 0524
jair.munoz-bugarin@moneyadviceservice.org.uk

Appendix

Over- and Under-indexed Ocean Variables

These tables show the proportion of individuals (survey respondents), split by over- and non-indebted, that meet each Ocean criteria.

E.g. 13% of respondents socially rent; 23% of those flagged as "over-indebted" socially rent.

The top fifty over-indexed variables are listed below. These are the characteristics you would expect to see in individuals that are most at risk of over-indebtedness.

Variable	Description	Total	Not Indebted	INDEBTED	Index (100=Avg)
lukp_a2_house2	Social renting	13%	11%	23%	197
hhdattrib_094	Multi-occupancy mixed sex household	2%	2%	3%	165
lukp_a2_age13	Aged 35-39	8%	7%	11%	155
lukp_a2_age11	Aged 25-29	5%	5%	8%	152
lukp_hhdval0to100k	Value of home: 0 to 100k	21%	19%	28%	145
lukp_a2_income1	Household income 0-9,999	9%	8%	11%	142
fr_de_nrse	Social grade E	8%	8%	11%	140
lukp_car0	Number of cars 0	20%	19%	26%	140
lukp_a2_house6	Flat or maisonette	16%	15%	21%	137
lukp_a2_children3	Children at home : 3+	6%	5%	7%	137
lukp_a2_age12	Aged 30-34	7%	7%	9%	137
lukp_a2_children4	Children at home aged 0-4	11%	10%	14%	137
lukp_occd13	Occupation: unemployed	3%	3%	4%	137
lukp_nbeds1	Number of beds : 1	8%	7%	10%	135
lukp_a2_children5	Children at home aged 5-10	13%	12%	17%	133
lukp_daily10	Daily Star	2%	2%	3%	132
lukp_a2_age14	Aged 40-44	9%	9%	11%	131
lukp_a2_hhd6	Household size : 5+ persons	7%	7%	9%	131
lukp_a2_age1	Aged 18-24	13%	12%	16%	130
lukp_a2_house3	Private renting	18%	17%	22%	130
fr_int_lifpen	Is likely to take out or switch supplier of life and pensions in next 12 months	4%	4%	6%	129
lukp_daily11	The Sun	14%	13%	17%	127
ips_sncomment	Often comments on friends pages	15%	14%	18%	126
hhdattrib_091	Single person household	30%	29%	36%	126
ips_snfollowed	Many Twitter followers	7%	6%	8%	126
ips_regwriteblog2	Blogs	6%	6%	7%	126
ips_regaccesseduc	Uses internet at school, college or university	9%	9%	11%	126

Variable	Description	Total	Not Indebted	INDEBTED	Index (100=Avg)
lukp_a2_children6	Children at home aged 11-15	13%	12%	15%	125
lukp_magwomengloss	Magazines read : womens glossy	7%	7%	9%	125
ips_snbrandlike	Likes, becomes fan of, or interacts with brand pages	16%	16%	20%	125
ips_regdate2	Online dating	5%	5%	6%	125
occ_a	Length of residence: 0-2 Years	9%	9%	11%	124
lukp_a2_age15	Aged 45-49	9%	9%	11%	124
ips_regaccesstvgame2	Uses internet on TV or games console	14%	13%	16%	124
ips_mobsocnet	Social networking on mobile	33%	32%	40%	124
lukp_sunday3	Sun Sunday	14%	13%	17%	124
fr_de_nrsd	Social grade D	16%	15%	19%	124
lukp_footteam	Interests : football supporter	9%	9%	11%	123
lukp_houseage7b	Home built 2006 or later	5%	5%	6%	122
occ_b	Length of residence: 3-5 Years	14%	14%	17%	122
lukp_occd10	Occupation: shop worker	7%	7%	9%	122
lukp_magmusicfilm	Magazines read : music and film	5%	4%	5%	122
fr_loa_2plus	Has 2+ loans	2%	2%	2%	122
fr_cur_basic	Has basic bank account	19%	18%	22%	122
ips_regaccesscafe2	Uses internet at library, cafe etc.	4%	4%	5%	121
lukp_a2_children2a	Children at home : 1	18%	18%	21%	121
ips_downldmusic	Downloads music	16%	15%	18%	121
ips_sharestuff	Shares content, e.g. video, articles or music	19%	19%	22%	121
ips_mobusesqr	Uses code reader (QR scanner) on mobile	11%	11%	13%	120
ips_useconsole	Uses games console	32%	31%	37%	120

The top fifty under-indexed variables are listed below. These are the characteristics you'd expect to see in individuals that are least at risk of over-indebtedness.

Variable	Description	Total	Not Indebted	INDEBTED	Index (100=Avg)
lukp_a2_age23	Aged 85+	0%	0%	0%	15
lukp_a2_age22	Aged 80-84	1%	1%	0%	22
lukp_a2_age21	Aged 75-79	2%	3%	1%	26
lukp_a2_age20	Aged 70-74	6%	6%	2%	28
lukp_a2_age19	Aged 65-69	10%	11%	3%	30
fr_inv_val6d	Has investments, value 100000+	1%	1%	0%	36
fr_inv_val6	Has investments, value 25000+	4%	4%	2%	42
fr_sav_val6d	Has savings, value 100000+	2%	2%	1%	42
fr_sav_val6	Has savings, value 25000+	9%	10%	4%	44
lukp_occd8	Occupation: retired	11%	12%	5%	46
fr_inv_bonds	Has investment bonds	2%	2%	1%	49
fr_inv_unittrust	Has Unit Trusts	1%	1%	1%	53
hhdattrib_093	Multi-occupancy single sex household	2%	2%	1%	56
lukp_maghome	Magazines read : home and gardening	3%	3%	2%	56
fr_inv_sharesisa	Has stocks and shares ISA	5%	5%	3%	56
lukp_house5	House owned outright	33%	35%	21%	60
lukp_interests55	Interests : antiques or fine art	7%	8%	5%	60
lukp_hhdval1mplus	Value of home: 1m plus	1%	1%	1%	60
lukp_sunday9	Sunday Telegraph	4%	4%	3%	61
lukp_holdest2	Holiday in Asia	2%	2%	1%	61
lukp_interests54	Interests : wine	3%	3%	2%	61
lukp_invhas	Has investments	13%	14%	9%	62
fr_de_nrsa	Social grade A	4%	4%	2%	62
fr_sav_val5	Has savings, value 10000-25000	9%	9%	6%	62
fr_inv_shares	Has stocks and shares	7%	7%	4%	62
lukp_a2_age18	Aged 60-64	10%	11%	7%	62
fr_inv_val5	Has investments, value 10000-25000	3%	3%	2%	63
lukp_daily6	Daily Telegraph	4%	4%	3%	63
lukp_a2_house5	Detached house	25%	27%	17%	63
lukp_hhdval750kto1m	Value of home: 750k to 1m	1%	1%	1%	63
lukp_hhdval500kp	Value of home: 500k plus	7%	8%	5%	64
lukp_magtravel	Magazines read : travel	3%	4%	2%	64
lukp_interests19	Interests : playing golf	6%	7%	4%	65
fr_sav_natsav	Has a National Savings product	17%	18%	11%	65
lukp_hhdval500to750k	Value of home: 500 to 750k	5%	5%	3%	65

Variable	Description	Total	Not Indebted	INDEBTED	Index (100=Avg)
fr_ins_struct2	Have home structure insurance	45%	48%	32%	67
lukp_interests32	Interests : self improvement / education	1%	1%	1%	67
lukp_a2_house4	Bungalow	8%	9%	6%	67
lukp_holtype7	Holiday : weekend short break	20%	20%	14%	69
lukp_interests11	Interests : environment / wildlife	7%	8%	5%	69
fr_mor_other	Has other mortgage (mixed, pension, PEP, Interest only, ISA etc.)	4%	4%	3%	70
fr_inv_val4	Has investments, value 2500-10000	3%	3%	2%	70
lukp_mortendow	Has endowment mortgage	1%	1%	1%	70
lukp_a2_income6c	Household income 100,000+	4%	4%	3%	70
lukp_nbeds4	Number of beds : 4	18%	19%	13%	70
lukp_paycredinfull	Always pays credit card balance in full	27%	29%	20%	70
fr_cds_spe500plus	Spent 500+ in last month on a credit card	15%	15%	11%	71
lukp_phlth	Has medical insurance (PMI) - pay all personally	3%	3%	2%	72
lukp_sunday10	Sunday Times	3%	3%	3%	72
lukp_holdest1	Holiday in Africa	3%	3%	2%	73

Model Variable References

Model Variable	Source of Data
Has loan for consolidation	FRS
Private renting	FRS
Social renting	FRS
Has 3+ children	FRS
Single parent	FRS
Northern Ireland	ONS
Social Grade D or E	FRS
Value of home <£100k, South East	DLG
Value of home <£100k, London	DLG
Unemployed, Wales & West Midlands	DLG
Household income <£10k, Household size 3+	FRS
Own home outright, Wales	FRS
Has savings £10k+	FRS
Aged 65-74	FRS
Aged 75+	FRS
Scotland	ONS

FRS = Modelled by CACI, based on data from the *Financial Research Survey,* GfK

DLG = Modelled by CACI, based on data from *Consumer Lifestyles, DLG*

ONS = Defined boundaries by the UK *Office of National Statistics*

the Money Advice Service

Money Advice Line **0800 138 7777***
Typetalk **1800 1 0300 500 5000**

If you would like this document in Braille, large print or audio format please contact us on the above numbers.

*Calls cost the same as a normal call, if your calls are free, it's included. To help us maintain and improve our service, we may record or monitor calls.

Information correct at time of publication (March 2016)

Money Advice Service
Holborn Centre
120 Holborn
London EC1N 2TD

March 2016

moneyadviceservice.org.uk