Developing a method to classify the adult social care workforce in England

Developing a machine learning method to identify and classify members of the adult social care workforce in England from Census 2021 write-in responses.

Contact:
Health Research Group and Data Science and Engineering Group
health.data@ons.gov.uk
+44 1329 444110

Release date:
31 March 2025

Next release:
To be announced

Table of contents

1 . Main points

- This article presents a method to identify and classify members of the adult social care workforce from free text responses to Census 2021 write-in questions.
- We analysed three linked data sources: Census 2021, the Care Quality Commission (CQC) Care Directory and Inter-Departmental Business Register (IDBR).
- We tested two different class sizes, two sentence embedders, an over-sampling technique, and implemented super learners (each comprising six models) with two voting styles (hard and soft).
- Super learner with soft voting, using a Doc2Vec embedder, and over-sampling on a larger dataset gave the strongest results, followed by Multi-layer Perceptron (MLP).
- It is more feasible to predict whether an individual works in adult social care and the type of service they work in than to predict their exact job, unless the job is well-defined.
- This research could be further developed through improved tools, an improved approach to labelling data, and exploring other data sources.

This method is experimental and therefore we do not report adult social care workforce population estimates. We welcome feedback on our approach and method.

2 . Overview

Understanding the size and composition of the adult social care workforce (ASC-WF) is a vital part of understanding supply and demand for the adult social care (ASC) sector. The Office for Statistics Regulation (OSR)'s 2020 Report on Adult Social Care statistics in England highlighted evidence gaps in the data landscape, including in ASC-WF. The Department of Health and Social Care subsequently commissioned the Office for National Statistics (ONS) to research how ONS data sources could be used to validate and build upon existing ASC-WF statistics.

The most robust statistics on the ASC-WF in England are reported by Skills for Care (SfC), the strategic ASC-WF development and planning body for England, in their The state of the adult social care sector and workforce in England report. Since 2020, SfC have produced their first accredited official statistics report, The workforce employed by adult social services departments in England, and have plans for developing their statistics, Our future plans and statistical governance (PDF, 320KB).

SfC achieve complete coverage of local authority care providers but lack complete coverage of the independent ASC-WF, which comprises 79.1% of filled posts, according to their estimates. The independent sector also includes providers not regulated by the Care Quality Commission (CQC). SfC have developed a robust Methodology for creating workforce estimates to weight their estimates for the independent sector, but exploration of ONS data sources could help validate and expand SfC statistics.

Census data collected by ONS have the advantage of near full population coverage. Census data also contain free text descriptions of individuals' employers and jobs, which can be used to train machine learning (ML) models to identify and classify people working in the ASC-WF. As well as triangulation, identifying the ASC-WF in census data would allow for more complex research questions to be explored, because of available characteristics in census data and existing linkage to administrative datasets.

Definition of the adult social care workforce

ONS developed a bespoke definition framework for this research, Definition framework for the adult social care workforce (Excel file, 73.3 KB). This is because Census 2021 was labelled using standardised frameworks, Standard Industrial Classification (SIC) and Standard Occupational Classification (SOC), which do not fully describe the ASC-WF. For example, SIC does not differentiate between adult and child social care or different types of non-residential social care, and several care-related jobs do not have specific SOC codes.

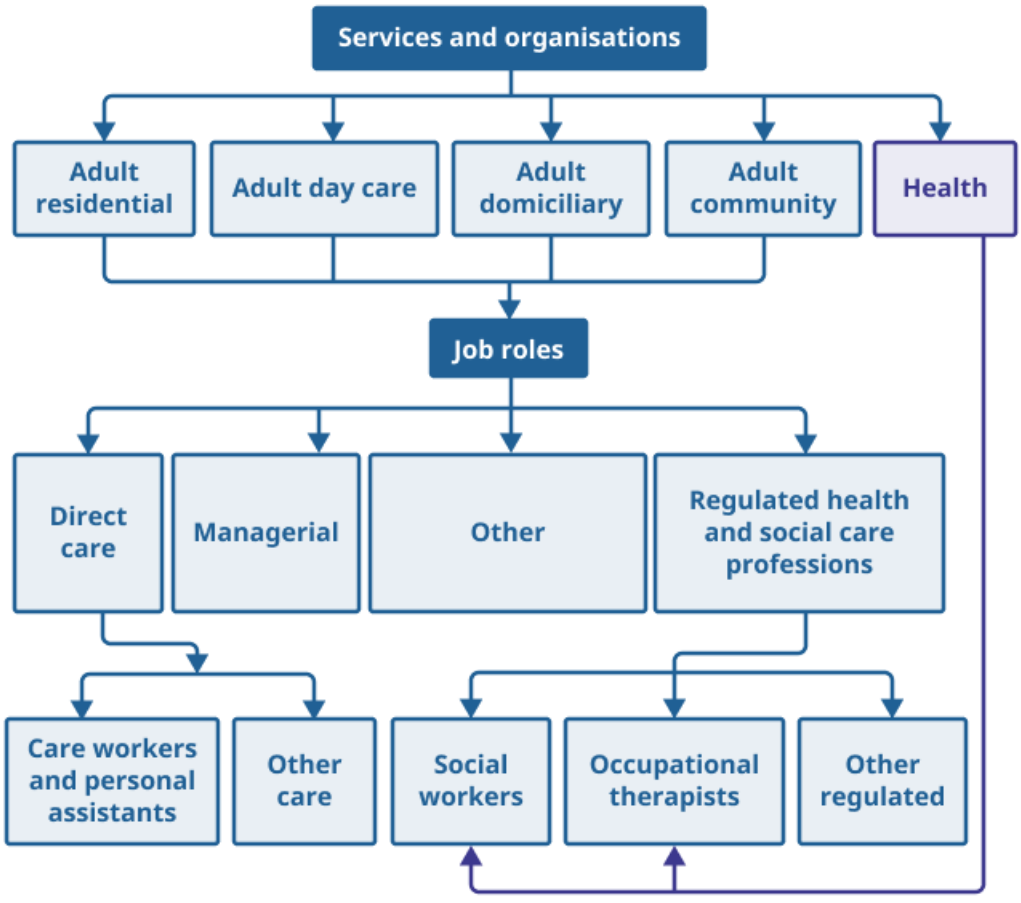
In this research, we define the ASC-WF as:

- the paid, employed workforce of organisations that provide adult social care, where the organisation has direct contact with care recipients
- all social workers and occupational therapists that work with adults regardless of employer

This research focuses on people employed by or as care providers. Therefore, our definition of the ASC-WF does not include voluntary or unpaid care provision, although these roles make a significant contribution to the social care system (for example, see Valuing Carers 2021 to 2022: the value of unpaid care in the UK (PDF, 1.28MB).

The full definition framework includes breakdowns by ownership type, service or organisation type, and job type. A simplified version of this framework was used for this research, comprising five broad categories of service or organisation type, four broad categories of job roles, and five more granular categories of job roles (Figure 1).

Figure 1: Hierarchy of service or organisation types and job roles in the adult social care workforce (ASC-WF) used in this research



Source: Office for National Statistics

Notes:

1. “Health” is only used as a category for social workers and occupational therapists not working for ASC organisations.

Aims and applications

The main purpose of this research was to develop an initial experimental method to identify and classify the ASC-WF from Census 2021 write-in responses, using supervised ML classification models, without only relying on SIC and SOC. Unsupervised ML methods were explored, however, use of a specific definitional framework, and prioritising interpretability, meant that unsupervised techniques were unsuitable. Modelling performance metrics are presented; however, because it is an experimental methodology, final ASC-WF population estimates are not included.

3 . Methodology

The overall method is summarised as follows:

1. Standardise responses, remove errors, and blank information-sparse columns from free text responses to Census 2021 employment questions.

2. Link Census 2021 to Care Quality Commission (CQC)'s Care Directory and the Inter-Departmental Business Register (IDBR), to provide information for data labelling (see step 4).

3. Filter the linked data to the study population and deduplicate so that one row represents one person.

4. Create a labelled dataset for training, testing and validating the classification models, using information from linked data and rules-based word-matching.

5. Conduct further data pre-processing (stop word removal, lemmatisation, and embedding) to convert free text into a modelling interpretable format.

6. Undertake stratified sampling to create train-test datasets.

7. Train six machine learning (ML) models and use these as part of super learner (SL) models to identify the most appropriate model for classifying the adult social care workforce (ASC-WF).

8. Interpret and evaluate ML model classifications.

The following sections cover these steps in more detail. See the "Strengths and limitations" subsection of Section 4: Results for further explanation of the rationale behind our approach.

Definitions of some of the technical terms in this methodology are available in Section 6: Glossary and Developing a method to classify the adult social care workforce in England (Excel file, 75.7 KB). All counts have been rounded to multiples of five to comply with statistical disclosure rules. All figures are reported to three decimal places.

We established a Methodological Advisory Group to provide input and feedback on our method, see the Collaboration sub-section of Section 5: Future developments for more information.

Data

Census

Every 10 years, there is a census of the entire UK population to provide a snapshot of individual and household demographics. The Office for National Statistics (ONS) is responsible for running the Census in England and Wales. For more information, see Quality and methodology information (QMI) for Census 2021.

To classify individuals as members of the adult social care workforce (ASC-WF), free text data from write-in responses to questions 41 to 44 from the Census 2021 provided the body of text for training models. These free text fields correspond to employer name, business description, job title, and job description. See the Census 2021 paper questionnaires for the question wording.

Inter-Departmental Business Register

The Inter-Departmental Business Register (IDBR) is the comprehensive list of UK businesses used by government for statistical purposes, providing the main sampling frame for business surveys, and is compiled from several administrative and survey data sources. For more information, see Inter-Departmental Business Register Data Quality. In this research, we used a quarterly snapshot of the IDBR taken on 16 March 2021, as the closest data to Census Day 2021 (21 March 2021).

Care Quality Commission

The Care Quality Commission (CQC)'s Care Directory is a publicly available dataset updated monthly using data obtained via the registration process of every "person" (an individual, partnership or organisation) who provides regulated care activity in England, as specified in The Health and Social Care Act (2008). It includes information about the care provider such as regulated activities provided, types of service provided, and provider user groups. ONS receives the Care Directory on a quarterly basis. Care Directory data for this study was from Quarter 1 (Jan to Mar) 2021 as the closest time to Census Day 2021 and the IDBR snapshot. However, some providers may have opened but not yet registered with the CQC by Census Day 2021.

The IDBR and Care Directory were used as sources for labelling service or organisation types (see "Labelling" subsection). See the "Strengths and limitations" subsection of Section 4: Results for further discussion of all data sources.

Data linkage

The three data sources were linked using deterministic methods, using a series of match-keys to pair records based on specific variables. The matching began with exact matches and progressed to allow for partial matches where necessary. The match-keys were based on variables including employer information (such as organisation name, address, and postcode), Companies House information, and Pay-As-You-Earn and Value Added Tax references.

One person in Census 2021 may link to multiple CQC or IDBR organisations for several reasons. For example, insufficient employment information may result in an individual record being linked to multiple entries on the Care Directory or IDBR. Additionally, one IDBR unit could have multiple rows with the same information, but different address naming conventions, and these "duplicate" records were retained to improve chances of finding a link.

Similarly, the Care Directory has multiple levels of organisation: brand, provider, and location. Where possible, records were matched on CQC location identity (ID), but otherwise they were matched on provider or brand IDs. This may result in one Census 2021 record linking to multiple CQC locations under one provider or brand.

The CQC was linked to the IDBR (CQC-IDBR linkage rate: 34.8% of CQC location IDs) and then Census 2021 was linked to CQC-IDBR (Census-CQC-IDBR linkage rate for working residents: 21.0%). Linkage rates are low because accuracy was prioritised over number of links, since the primary purpose of the linkage was to provide information for labelling (see the "Strengths and limitations" subsection of Section 4: Results for further information). This resulted in a total of 44,404,440 links, and 8,901,815 unique Census 2021 records linked. Census 2021 records which did not link to CQC or IDBR were retained in our dataset.

For each match-key, a sample of links were manually reviewed to estimate the number of true positives (correct links) and false positives (incorrect links). This allows calculation of weighted estimates of precision for match-keys. Precision is the number of true positives divided by the sum of true positives and false positives. This was 96.1% without uncertain links, and 37.6% for uncertain links.

For further information on the linkage method and quality, including discussion of the linkage rate and reviewing rejected records to estimate recall, see Care Quality Commission and Inter-Departmental Business Register linkage to Census 2021.

Data cleaning

Data cleaning was required to standardise the write-in responses to Census 2021 and remove any obvious errors to aid rule-based word matching. For example, invalid and information-sparse responses (columns with one or fewer characters) were blanked, non-alphanumeric characters (such as punctuation) were removed, digits that should be characters were replaced (for example, "people" spelled with a zero instead of the "o" to "people" spelled correctly), and formatting was standardised (such as converting to lower case, deleting extra blank spaces). Where a response to a question was "as above" or similar, this response was blanked.

Filtering and deduplication

The cleaned dataset was filtered to the study population of interest: usual residents with a workplace address in England who were employed at time of Census 2021 (25,536,480 people). The filtered Census 2021 data were then joined via the linkage file to the CQC-IDBR (8,435,760 linked Census 2021 IDs).

The linked dataset was deduplicated to one record per person as follows:

1. Matched records where the match-key precision was less than 90% (32,634,690) had CQC and IDBR information removed and were treated as non-matches.

2. Remaining matched records where a person linked to multiple CQC location IDs (153,585) were flagged, had CQC information removed, and were treated as non-matches.

3. Remaining matched records where a person matched with IDBR rows with multiple different Standard Industrial Classification (SICs) codes (5,311,215) had IDBR information removed and were treated as non-matches.

4. All remaining matched records were deduplicated by Census 2021 ID and match-key.

Finally, Census 2021 IDs that were missing a response in all four columns, including blanked responses (254,810), were dropped because these data cannot be used for modelling. This resulted in a cleaned, deduplicated dataset of 25,281,670 records. Of these records, 4,770,630 were linked to the CQC and, or the IDBR; 436,790 records were linked to a single CQC location and 4,610,075 had information from IDBR.

Labelling

In the absence of trained occupational coders, we undertook a multi-step approach to create a labelled dataset for training, testing, and validating the classification model. Where records successfully linked to a single CQC location, the record was assigned to a class based on information from the Care Directory. For example, if the record had a CQC service type of "care home service with nursing", they would be labelled as "adult residential".

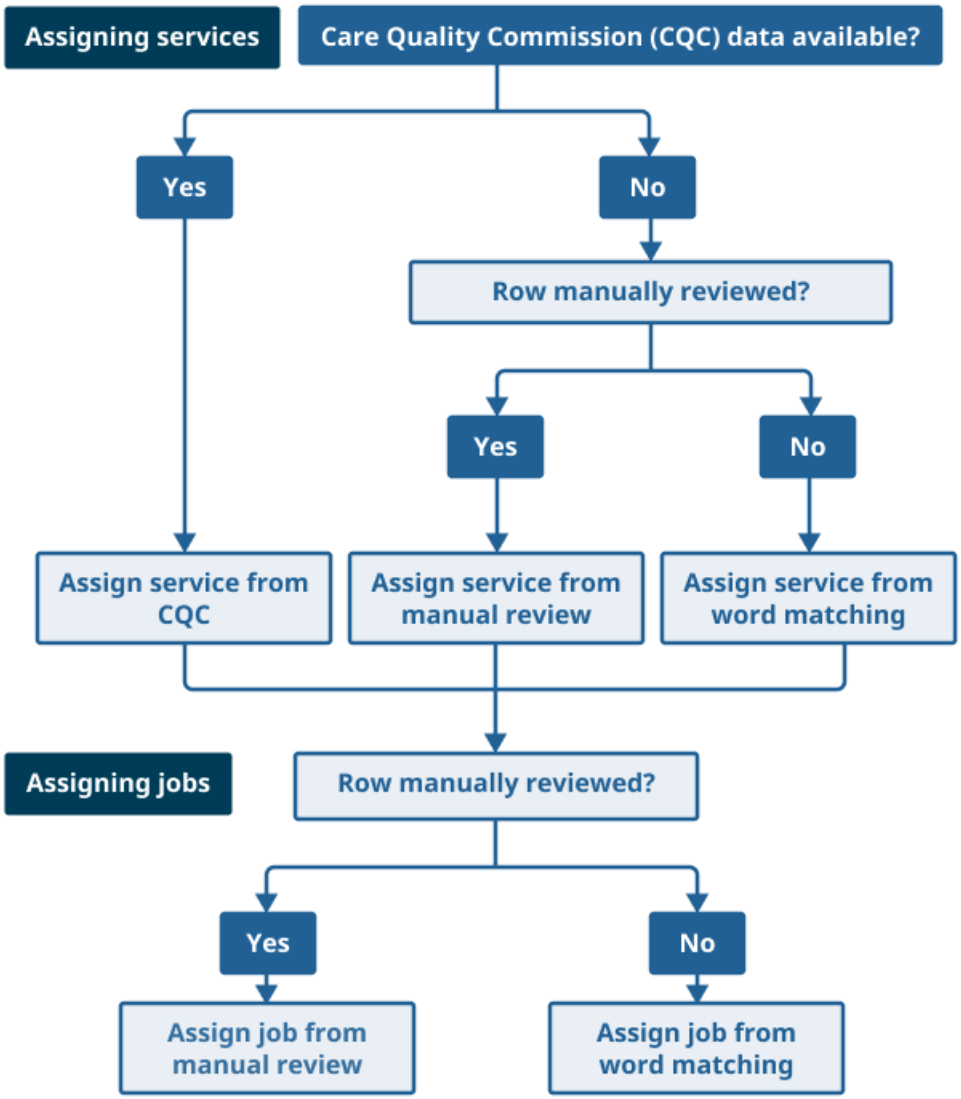
A different labelling strategy was required for:

- Census 2021 records that linked to multiple CQC location IDs
- Census 2021 records that did not link to the Care Directory to a sufficient degree of precision, including both missed matches and genuine non-matches
- assigning jobs, because job information is not included in the Care Directory

For these, records with SIC and Standard Occupational Classification (SOC) codes (from Census 2021 and IDBR) were searched for key words and phrases related to ASC in the write-in responses (for example, a respondent stating they work for a care home). Any rows that contained sufficient information were flagged as potential ASC-WF. These flags, along with the CQC labels, were then used to create variables for service or organisation type and job roles, corresponding to the definitional framework.

A small number (295) of rows labelled as "other community" were manually reviewed, because it is a less well-defined category, with fewer unique associated key words. Two coders independently reviewed each row and reconciled disagreements through discussion to assign labels according to the definitional framework. Labels from manual review were prioritised over word-matching labels. Figure 2 summarises the multi-stage approach and prioritisation of different sources of information for labelling the data.

Figure 2: Flowchart showing the multi-stage approach to creating a labelled dataset for training, testing, and validating models



Source: Office for National Statistics

Notes:

1. Social workers and occupational therapists are included regardless of service or organisation type, so for these occupations, jobs are assigned before services.

Further data pre-processing

Prior to creating the train-test datasets, additional data processing was conducted to prepare the data for analysis. We merged response columns into one column containing all free text, removed stop-words (common words which are not informative such as "the", "is", "and"), and implemented lemmatisation (simplifying the variety of words by reducing them down to a shared core word; for example, "improved" and "improving" become "improve").

We tested two embedding techniques to evaluate their impact on the modelling. Embedders convert free text to vector (numerical) representations. Multiple methods were explored (see the "Strengths and limitations" subsection of Section 4: Results), but the chosen embedders were Text Frequency- Inverse Document Frequency (TF-IDF) and Doc2Vec.

Creating train-test and validation datasets

To ensure only clear examples of ASC-WF and non-ASC-WF were used in training, validation and testing, we excluded records which had some indicators of ASC-WF but were not sufficient for a label (for example, if it was not clear if the person was working in social care for adults or children, or the ASC-WF job not sufficiently described).

Before creating the train-test and validation datasets, we reduced the sample size of the entire dataset to ensure the machine learning (ML) models run without encountering computational limitations. We applied a down-sampling technique to set a maximum sample size for the classes, preserving the information of minority classes while reducing the sample size of majority classes. While this method mitigates class imbalance and potential loss of information for minority classes, it results in some loss of information for majority classes.

Only records with unique merged responses were retained when creating train-test and validation data, to avoid data leakage, where the same processed response is present in both train and test data. This would compromise the unseen nature of the test data, important for obtaining valid and meaningful model performance and metrics.

At this stage, the labelled dataset was ready to be split into train-test-validation datasets for the binary target label of ASC-WF versus not ASC-WF. The proportions of this split were 80,10,10, respectively. The train dataset comprised 88,235 rows, of which 54,205 (61.4%) were labelled ASC-WF. The test and validation datasets each comprised 11,030 rows of which 6,775 (61.4%) were labelled ASC-WF.

Those defined as ASC-WF by the binary classifier were then split into further train-test-validation datasets, also with 80,10,10 splits, for these three target variables:

- service or organisation type ("service")
- job roles ("job")
- granular job

To ensure a robust training and evaluation process, a hierarchical train-test split was implemented, ensuring the three subsequent splits were based on training data from the first hierarchical layer, minimising the risk of data leakage and preserving the integrity of model performance. Additionally, these splits took a stratified approach in conjunction with a random state, to ensure consistent and representative sampling across all target variables. The aim was to model the organisational structure of the ASC-WF represented in Figure 1.

Following the split, an imbalance in class distribution was identified, affecting the "adult day care" category and other job-related classes. To overcome this, we applied Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) features, which generates synthetic data while considering categorical (target variables) and numerical features (text embeddings).

Modelling

To train the ML models, and find the optimal settings, we tuned the hyperparameters with Bayesian search optimisation, in conjunction with repeated stratified K-Fold cross validation. These techniques combined provide a comprehensive, yet computationally efficient method in finding the optimal hyperparameters, guided by which combination produces the best scoring model on left-out data.

To evaluate the models' fit to the data, we compared the results of the hyperparameter tuning using the best score metric with a validation score. The two scores were similar for all the models, showing that overfitting has been limited in the training of the models.

We explored six different supervised ML models in our analysis:

- Random Forest: Random Forest classifier (scikit-learn 1.6.1)
- Logistic Regression: Logistic Regression (scikit-learn 1.6.1)
- K-Nearest Neighbours Vote: Nearest Neighbours classifier (scikit-learn 1.6.1)
- Multi-layer Perceptron (MLP): MLP classifier (scikit-learn 1.6.1)
- Stochastic Gradient Descent: SGD classifier (scikit-learn 1.6.1)
- XGBoost: XGB classifier (xgboost 3.0.0)

To overcome strengths and weaknesses of different models, we employed a super learner (SL) approach. SLs work by pooling results from individual models. Each model's prediction is considered as a vote toward a meta-prediction, and we considered two voting styles: hard and soft (see Developing a method to classify the adult social care workforce in England (Excel file, 75.7KB)).

All modelling was conducted using Python in a cloud-based environment using Central Processing Unit (CPU)-driven techniques and 64 GiB memory.

4 . Results

This research explores the feasibility of classifying the adult social care workforce (ASC-WF). As stated, we are not presenting ASC-WF population estimates because this research is at an early stage of methodological development. To address our hierarchical multi-label classification problem, we explored various embedding and machine learning techniques. We showcase the predictive power of different combinations of techniques and compare their performance at each hierarchical layer (binary, service, job, granular job). We discuss potential methodological improvements and research in Section 5: Future developments.

Model accuracy metrics are available in Developing a method to classify the adult social care workforce in England (Excel file, 75.7KB).

We employed hierarchical classification, prioritising predicting the top layer (binary ASC-WF classification) because this affects all subsequent layers. When using binary classification to predict whether a class is ASC-WF or not, the evaluation metrics differ because of the conceptual context. False positives would be more costly at this layer because records misclassified as ASC-WF would affect the downstream models, hence Area Under the Receiver Operating Characteristic (AUC-ROC) was prioritised for this classification. For layers below the binary layer (service, job, granular job), we prioritised F1 to balance precision and recall.

First, we considered accuracy, which measures the percentage of correct predictions using our test dataset. While intuitive to understand, it is less appropriate in cases of class imbalance, such as in this research (for example, "adult day care").

In Figure 3, we present the overall accuracy for each model (the proportion of correct predictions using the test dataset), by each layer of the hierarchy, using Doc2Vec embedding with a sample size of 54,205. As expected, the overall accuracy decreases as we move down the hierarchy, because of:

- the decreased sample size below the first binary layer
- less well-defined classes such as "other"
- the increasing number of classes

The model with the highest accuracy for binary ASC-WF classification and job classification was the Multi-layer Perceptron (MLP) model (accuracy of 0.918 and 0.849 respectively). Yet, for service and granular job, the super learner (SL) with soft voting provides the best performance (accuracy of 0.860 and 0.794 respectively). Nevertheless, at each level, multiple models had similar accuracy scores, increasing our confidence in the underlying data.

Figure 3: Overall accuracy for each adult social care workforce (ASC-WF) classification model by each layer of ASC-WF hierarchy

Notes:

1. Maximum sample size per class was 20,000.
2. Doc2Vec embedding was used and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) has been applied.

Download the data

Downloac

Next, we calculated AUC-ROC scores to evaluate the models' ability to distinguish between classes. AUC-ROC scores consider both true positives (sensitivity) and false positives (specificity), making them particularly useful for an imbalanced dataset.

Figure 4 presents the AUC-ROC scores for each hierarchical layer using Doc2Vec with a sample size of 54,205. The adult social care (ASC) binary classification maintains high scores, with MLP as the best-performing model (AUC-ROC score of 0.966), followed by SL with soft voting (0.965). For multi-class (not binary) classifications, AUC-ROC scores have decreased in all models. The Logistic Regression model achieved the highest AUC-ROC score of 0.646 for service, however, the SL model using soft voting would be preferred for classification because of its superior accuracy (0.860, Figure 3). When classifying job and granular job, all models have AUC-ROC of 0.514 or below, indicating the models are behaving on par with or worse than random classification and therefore would not be recommended for use.

Figure 4: Area Under the Curve – Receiver Operating Characteristic (AUC-ROC) evaluation of machine learning classification models

Notes:

1. Maximum sample size per class was 20,000.
2. Doc2Vec embedding was used and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) has been applied.
3. Super learners with hard voting do not offer probabilities for predictions, so AUC-ROC scores cannot be derived.

Download the data

Downloac

For service, job, and granular job classification models, we used F1 score as an additional means of gauging model performance because of its balance of recall and precision. Figure 5 shows that SL with soft voting had the highest F1 score across all service classes.

Figure 5: F1 evaluation of adult social care workforce service machine learning classification models

Notes:

1. Maximum sample size per class was 20,000.
2. Doc2Vec embedding was used and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) has been applied.

Download the data

Downloac

Figure 6 shows the F1 scores for job classification. In well-defined categories like “direct care” and “regulated health and social care professions”, MLP scores highest at 0.894 and 0.875 respectively, closely followed by SL with soft voting. However, in the less well-defined categories, SL with soft voting slightly outperforms MLP, scoring 0.786 for “managerial” and 0.716 for “other”.

Figure 6: F1 evaluation of adult social care workforce job machine learning classification models

Notes:

1. Maximum sample size per class was 20,000.
2. Doc2Vec embedding was used and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) has been applied.

Download the data

Downloac

At the most granular job level (Figure 7), SL with soft voting is the top performer overall, achieving the highest F1 scores across all categories, with different models coming in second place for different classes. Notably, SL with soft voting had a greater effect compared with the second-place models for less well-defined categories (for example, 0.561 versus 0.515 for XGBoost in "other care providing"), compared with more well-defined categories (for example, 0.843 versus 0.833 for XGBoost in "care workers"). This indicates that the nuanced aggregation of meta predictions by the SL with soft voting is particularly beneficial for less well-defined categories. In addition, most models perform around or below chance for “other care providing” and “other regulated health and social care professions”, suggesting it is these less well-defined classes having a detrimental effect on the AUC-ROC scores for these models (Figure 4).

Figure 7: F1 evaluation of adult social care workforce granular job machine learning classification models

Notes:

1. Maximum sample size per class was 20,000.

2. Doc2Vec embedding was used and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) has been applied.

3. The "care workers" classification also includes personal assistants.

Download the data

Downloac

Embedders

In addition to Doc2Vec, we tested Text Frequency-Inverse Document Frequency (TF-IDF) as an embedder, and model outputs can be found in Developing a method to classify the adult social care workforce in England (Excel file, 75.7KB). We tested maximum class sizes of 2,000 and 20,000 for Doc2Vec, but only a maximum class size of 2,000 for TF-IDF, because its high dimensionality limited the size of data that could be modelled. In terms of overall accuracy, models using TF-IDF embeddings generally outperformed those using Doc2Vec. For example, for the binary ASC-WF classification model, the SL with soft voting was the highest performing model for both, with an accuracy of 0.891 for Doc2Vec and 0.979 for TF-IDF. However, the differences in AUC-ROC scores were smaller and sometimes in favour of Doc2Vec.

Pre-trained models such as Doc2Vec offer several advantages, including dimensionality control based on user input, allowing for consistent embedding sizes. Set dimensions enable training with larger datasets, providing more examples and potential variety. Doc2Vec's method enables it to understand the semantic meaning of words and their differences, and as sample size increases, so too does the accuracy (for example, from 0.891 to 0.912 for SL with soft voting for the binary classifier). However, it does not fully understand context, which can be a limitation (for example, "bank account" versus "river bank"). Including contextual understanding in embeddings could be beneficial, especially for specific areas of interest like nursing.

In contrast, TF-IDF, being an unsupervised model, derives its dimensionality from the unique occurrences of each word in the available text, making it more suitable for smaller samples. The simplicity of TF-IDF provides transparency, as each dimension is based on the presence and weighting of words in the body of text, making feature importance and decision-making explanations more accessible. However, TF-IDF treats spelling errors, non-words, and superfluous information equally, requiring further preprocessing to address, and in turn, reduce dimensionality. The importance of certain acronyms (for example NHS (National Health Service) and OT (Occupational therapist)) limits the capacity to remove short words or creates the need for an exhaustive dictionary of relevant acronyms. There could also be limitations from single acronyms having multiple context-dependent meanings.

Strengths and limitations

Strengths

The Definition framework for the adult social care workforce (Excel file, 73.3KB) for the adult social care workforce (ASC-WF) has advantages over Standardised Industrial Classification (SIC) and Standard Occupational Classification (SOC) frameworks because of its relevance and granularity (for example, distinguishing "adult domiciliary" from "adult community" and "adult day care"). This research also incorporates information from non-Care Quality Commission (CQC)-regulated sectors ("adult day care" and most "adult community" services), for which Skills for Care do not have complete coverage.

Another strength is the near full population coverage from Census 2021, and the inclusion of independent sources of information (CQC's Care Directory and Inter-Departmental Business Register (IDBR)) to assign labels to the train-test datasets.

Our comprehensive approach of testing several models, including a super learner, was an important strength. Super learners consider outputs from various models and are not reliant on the fitness-for-purpose of any individual model. Finally, exploration of hard and soft voting styles also strengthened our approach, by allowing our evaluation metrics to be maximised.

Limitations

Firstly, there was incomplete linkage between Census 2021, CQC Care Directory, and IDBR (see Care Quality Commission and Inter-Departmental Business Register linkage to Census 2021), meaning that some information could be missed. The low linkage rate is because we prioritised accuracy over number of links to ensure correct labelling when creating train-test and validation data. In addition, not all care providers will be represented on the IDBR, such as councils. Similarly, only providers of regulated activities need to register with CQC, and there may be a lag between a provider opening or closing and registering or de-registering with CQC. Therefore, several providers will not be present in the Care Directory.

Secondly, each dataset has some underlying issues. Some of these could be solved in the future, however, other issues are inherent limitations of the data. There is considerable variation in the quality and quantity of information provided by respondents in Census 2021. In the final labelled data, 87.7% of respondents had completed all four questions with at least two characters per column. Those with full or partial responses may not have provided sufficient information for labelling, even by human occupation coders.

We selected our study population based on workplace address because individuals may live in England but work elsewhere in the UK, or vice versa. However, Census 2021 was conducted during a period of lockdown because of the coronavirus (COVID-19) pandemic, which may affect how people report their place of work. His Majesty's Revenue and Customs (HMRC) Coronavirus Job Retention Scheme statistics indicate 3,616,500 people in England were on furlough on 2021 Census Day.

Thirdly, we could not address all data cleaning issues, such as using abbreviations and acronyms, people working multiple jobs, detecting and filtering responses in languages other than English, and spelling errors. Such responses remain in the data and may affect the accuracy of the word matching code, and therefore the training dataset and model.

Regarding methodological limitations, to obtain labelled data for supervised learning, our preferred approach would be to manually label a representative sample covering all possible combinations of services and jobs within the Definition framework for the adult social care workforce (Excel file, 73.3KB). Each record would be independently coded by separate people trained in occupational coding. However, this process is very resource intensive and was not a feasible option for this project. Therefore, an alternative multi-step approach was implemented (see the "Labelling" subsection of Section 3: Methodology), and each information source has strengths and limitations.

Information from CQC:

- is accurate as information comes from the regulator and is independent to free text data
- is not reliant on census respondents providing sufficient information
- has limited coverage, as CQC does not regulate "adult day care" or most "adult community" services

Information from word matching:

- is large scale
- has potential inaccuracies as it is rule-based, relying on exact matches to key words and phrases
- is not independent to the free text, introducing risk of circularity, although this is mitigated by incorporating SIC and SOC codes

Information from manual review:

- is accurate because humans are coding against a detailed framework
- has a high resource requirement leading to reduced scale

Finally, some preferred tools for embedding and modelling were not available for this research. We tested TF-IDF and Doc2Vec and have outlined their strengths and limitations.

Text Frequency-Inverse Document Frequency (TF-IDF):

- is simple, quick to run, easy to interpret, and works well for sparse and high-dimensional data
- is rigid and does not capture the meaning of words or their relationships
- is dependent on the input body of text
- fails to consider word order

Doc2Vec:

- provides some understanding of relationships between words and offers more tuning parameters than TF-IDF
- demonstrates learnt bias
- may assign negative values (minus 1 to positive 1) so is not natively compatible with models that require only positive values
- has no understanding of nuance or context
- is computationally expensive
- requires substantial data for effective training

Summary

We have demonstrated the predictive power of different models across each layer of the hierarchical classification. The results of this feasibility research indicate that we can predict the overall population of the ASC-WF with a strong AUC-ROC score of 0.965 and an accuracy of 0.918.

Considering a holistic view of evaluation metrics, the SL with soft voting was predominately the best-performing model for every hierarchical classification level, closely followed by MLP. Of the two embedders, the greater linguistic understanding of Doc2Vec, and the flexibility it offers in setting dimensions, makes it a more practical choice for modelling with large datasets such as census data.

We can predict type of service using SL with soft voting, with an AUC-ROC of 0.635 and overall accuracy of 0.860. However, further research is needed for more granular job layers. As observed with the F1 and AUC-ROC scores, there was a greater decline in predictive confidence deeper in the hierarchy, where more categories are less defined.

5 . Future developments

This release presents feasibility research, and we are not currently planning to produce population estimates of the adult social care workforce (ASC-WF). Several aspects of the method would need development before such estimates could be produced, such as a fully labelled dataset coded by pairs of experts in occupational coding.

Potential pre-processing developments include exploring more sophisticated spell-checking tools, filtering responses in languages other than English, and more specialised embedding tools. We could also explore the optimal dimensionality of embedders, which were produced with 100 dimensions (for Doc2Vec) in the current research. Exploring feature importance with tools such as SHAP (SHapley Additive exPlanations) could aid in reducing data dimensions and improve understanding of model behaviour.

We considered Sentence-BERT, a pre-trained large language model that has more contextual embedding. It would have strengths over Text Frequency-Inverse Document Frequency (TF-IDF) and Doc2Vec because it understands the context and nuance of words, produces embeddings suitable for downstream natural language processing tasks, and is more effective for tasks requiring an understanding of sentence-level relationships. However, it was not available to us at the time of analysis.

Similarly, more sophisticated modelling techniques such as linear neural networks should be explored. In addition, a new tool, recently developed by the Office for National Statistics (ONS) Data Science Campus, may be useful if adapted to our definitional framework: ClassifAI: Exploring the use of Large Language Models (LLMs) to assign free text to commonly used classifications.

Our research focused on person-level predictions rather than organisation-level predictions; linking the census to Inter-Departmental Business Register (IDBR) and Care Quality Commission (CQC) allows for analysis at an organisation level. However, this would require more complete linkage of the IDBR to census and CQC, and refinement of our handling of one-to-many, many-to-one, and many-to-many links. Finally, the production of full ASC-WF population estimates would require weighting, to account for incomplete linkage, and rows dropped because of incomplete census information or imputed responses.

Following methodological development, full ASC-WF population estimates could be produced with more granular breakdowns of service and job type, breakdowns by ownership (public, private, or non-profit), and breakdowns by characteristics. This would allow for more nuance in understanding the size and structure of the ASC-WF, for example, distinguishing "personal assistants" from "care workers" as they have distinct characteristics.

The definition framework and models described here could also be applied to other data sources, such as surveys, or job advert data supplied to ONS by Textkernel (for an example, see Labour demand volumes by Standard Occupation Classification (SOC 2020), UK). This would allow timelier estimates of ASC-WF characteristics and labour demand, respectively. More complex research questions could be explored by linking census or survey data to other administrative data, such as income and health data. Finally, future research should consider comparison and harmonisation with the four nations.

Feedback

We welcome any feedback on this methodology to the ONS Health Data inbox: Health.Data@ons.gov.uk.

Collaboration

This work was commissioned by and developed in collaboration with the Department of Health and Social Care. We also worked with Skills for Care, who advised on the definition and labelling approach, and reviewed this paper ahead of release.

We established a Methodological Advisory Group of experts who advised throughout on the methodological development, including peer review of code and reviewing the paper ahead of publication, and would like to thank them for their contribution:

- Skills and Human Development and Household Resilience teams at ONS
- Centre for Care, University of Oxford
- Digital Health Group, Department of Population Health Sciences, King's College London
- Social Care Wales

6 . Glossary

Embedders

Techniques that convert free text to vector (numerical) representations. We tested two embedders for this research:

- Term Frequency-Inverse Document Frequency (TF-IDF)
- Doc2Vec

Evaluation metrics

Used to assess the performance of the models. The higher the score (0 to 1) the more well-balanced the predictive capabilities of the model. Scores of 0.5 indicate the model is performing at chance. We assessed the following metrics:

- accuracy
- Area Under the Curve-Receiver Operating Characteristics (AUC-ROC)
- precision
- recall
- F1

Machine learning classifiers

Used to categorise data into predefined classes or labels, machine learning (ML) classifiers are a type of supervised learning model trained on a labelled dataset, where each data point is associated with a specific class. The classifier learns the patterns and relationships and uses this knowledge to predict the class of new, unseen data points. We tested the following ML classifiers, employing a super learner approach:

- Random Forest
- Logistic Regression
- K-Nearest Neighbours Vote
- Multi-layer Perceptron
- XGBoost
- Stochastic Gradient Descent

Train (training) dataset

A train dataset is a representative sample of the data where each row has been labelled with the correct classification, providing examples for the ML model to learn the types of responses associated with each classification.

Test dataset

A test dataset is an independent labelled dataset that is used to test how well the model has learned from the train dataset.

Validation dataset

The final validation or holdout dataset is also labelled but is never used in the train or test stages and is used to assess the fit of the model.

7 . Related links

Care Quality Commission and Inter-Departmental Business Register linkage to Census 2021
Methodology | Released 5 March 2025
Linkage methodology and quality information for Care Quality Commission and Inter-Departmental Business Register linkage to Census 2021, to facilitate research into estimating the size and structure of the adult social care workforce.

The state of the adult social care sector and workforce in England
Skills for Care report | Published October 2024
This report provides a comprehensive analysis of the adult social care workforce in England and the characteristics of the 1.59 million people working in it.

Identifying different roles in the social care sector using online job advertisements
Data Science Campus blog | Published 10 November 2022
Blog published by the Office for National Statistics (ONS) Data Science Campus outlining a method to identify adult social care jobs from online job adverts.

8 . Cite this working paper

Office for National Statistics (ONS), published 31 March 2025, ONS website, methodology, Developing a method to classify the adult social care workforce in England.